UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ENGENHARIA GEOGRÁFICA, GEOFÍSICA E ENERGIA

Segmentação socioeconómica das freguesias de Portugal Continental

José Maria Mota Dias Miranda Alves

Mestrado em Sistemas de Informação Geográfica – Tecnologias e Aplicações

Trabalho de projeto orientado por:
Professora Doutora Cristina Maria Sousa Catita
Doutora Rita Matildes

2018

Agradecimentos

Quero dedicar este projeto à minha família, aos meus Pais e às minhas duas irmãs, por todo o apoio, paciência e compreensão. Sem vocês, nada disto teria sido possível.

Queria agradecer a todos os meus amigos, em especial ao R. Pirez, P. Antunes, K.M. Ramalho, D.D. Coelho, J.T. Rodrigues por tudo o que proporcionam à minha vida diariamente.

Queria agradecer à professora Cristina Catita, não só pela influência que teve durante o desenvolvimento deste projeto, mas também por toda orientação, apoio e paciência demonstrado durante a realização deste mestrado. Queria também agradecer a Doutora Rita Matildes pela orientação e apoio disponibilizado para o desenvolvimento deste projeto.

Resumo

Com o aumento da competitividade no setor retalhista, proveniente do avanço tecnológico, cada vez se torna mais importante conseguir localizar potenciais clientes e conhecer as suas tendências de maneira a conseguir fornecer o serviço adequado às necessidades dos mesmos. Isto torna-se possível devido a técnicas e metodologias desenvolvidas para criar bases descritivas da população, permitindo a entidades caracterizar os seus clientes e assim otimizar e desenvolver produtos que satisfaçam a procura dos mesmos.

O objetivo deste projeto foi encontrar uma metodologia que permitisse criar uma segmentação socioeconómica das freguesias de Portugal Continental, e que sirva de ferramenta para o auxílio à tomada de decisão no domínio do retalho utilizando dados proveniente dos censos e das bases de dados da *ESRI*. Esta informação foi particionada utilizando o algoritmo *k-means*, resultando em 5 *clusters*, tendo sido proposta uma descrição para cada um e, essa descrição contempla um modelo que incorpora um conjunto de variáveis socioeconómicas que foram analisadas individualmente de forma a melhor se compreender a sua distribuição em cada *cluster*. Por fim, a solução gerada através desta metodologia foi comparada com um subconjunto de dados demonstrativos de consumo de clientes de forma a validar a sua eficácia.

Palavras-chave: Geodemografia, segmentação, agrupamento, *k-mean*s, índice de autocorrelação espacial.

Abstract

With the increase of competitiveness in the retail sector due to technological advance, it is ever important to be able to locate potential clients and to know their tendencies in order to provide the best suited service to their needs. This became possible due to techniques and methodologies developed to create population descriptive bases, allowing entities to characterize their clients and thus optimize and develop products that meet their demand.

The goal of this project is to find a methodology to create a socioeconomic segmentation of the parishes of Portugal Continental, as a tool to improve decision making in the retail sector, using data from the census and data from *ESRI* databases. This information was partitioned using the *k-means* algorithm, resulting in 5 *clusters*, and a description was proposed for each one. This description contemplates a model that incorporates a set of socioeconomic variables that were analyzed individually in order to better understand their influence on each *cluster*. Lastly, the generated solution was compared to a subset of data demonstrating consumption data to validate its efficiency.

Keywords: geodemography, segmentation, clustering, k-means, spatial autocorrelation indicators

Índice

Índice de figuras

Índice de tabelas

1 – Introdução

1.1. Enquadramento do tema

A segmentação de mercado é um elemento fulcral dentro do *marketing* quando aplicado a países industrializados. No início deste século, o desenvolvimento industrial de vários setores económicos induziu estratégias de produção em massa, orientadas ao fabricante e focadas na redução de custos de produção em vez da satisfação dos consumidores. Mas a evolução dos processos de produção e a afluência de consumidores levou à diversificação da procura, o que potenciou uma vantagem competitiva das entidades que identificavam as necessidades de grupos de consumidores, devido à capacidade de desenvolver a oferta correta para um ou mais sub-mercados. A segmentação de mercado foi primeiramente introduzida por Smith (1956), definindo o conceito da seguinte maneira: "Segmentação de mercado envolve ver um mercado heterogéneo como vários pequenos mercados homogéneos, em resposta a diferentes preferências, atribuídas ao desejo dos consumidores de uma satisfação mais precisa das suas variadas necessidades.".

Através da segmentação torna-se possível direcionar campanhas de *marketing* a determinados alvos de interesse, pois com a inclusão de diferentes indicadores socioeconómicos irá ser possível uma caracterização mais individual do consumidor. Estas caracterizações permitem também a adaptação de produtos às necessidades de um ou mais segmentos, identificar novos nichos de mercado, otimizar a localização de novas lojas e produtos comercializados, entre outros fatores que serão preponderantes na construção de vantagens competitivas e na maximização de lucro.

Com o desenvolvimento dos sistemas de informação, os analistas de mercado começaram a ter acesso a informação enriquecida sobre o comportamento real dos consumidores através de bases de dados espaciais e segmentações geodemográficas. Assim, a extração de informação espacial ganhou uma elevada importância na abordagem de segmentação de clientes baseada nas suas diferentes características e hábitos, o que provocou o desenvolvimento das técnicas de particionamento de dados. Este desenvolvimento é importante, não só na vertente do *GeoMarketing*, pois o conhecimento da distribuição espaciotemporal da população a um nível local permite melhorar o planeamento territorial, a gestão de riscos, estudos ambientais e da saúde, entre outros. (Freire, 2010).

O *GeoMarketing* é uma vertente que combina a capacidade de visualização e analise geográfica com técnicas e informação derivada do *marketing*, de maneira a aumentar a venda de produtos, serviços ou ideias (Freire, 2010). O aparecimento desta vertente foi facilitado pelos avanços na análise espacial e visualização dos Sistemas de Informação Geográfica (SIG), apesar da sua utilidade para a realização de estudos económicos continua por ser explorada (Cheng et al., 2007; Mishra, 2009).

Como ferramenta de apoio à tomada de decisão, as análises de *GeoMarketing*, podem ser divididas em etapas sequenciais: (1) formular o problema, (2) obter e processar os dados disponíveis e necessários, (3) efetuar a análise e (4) definir conclusões e recomendações. Para a caracterização da oferta (localização, serviços, produtos, concorrentes) e da procura (população, clientes existentes e potenciais) é necessário analisar conjuntos de dados espaciais e não espaciais (Freire, 2010).

1.2.Objetivo do projeto

O objetivo deste projeto é propor uma segmentação das freguesias de Portugal Continental de acordo com as características socioeconómicas das populações que nelas habitam, tendo em conta as distribuições de idades, gênero, condições de habitação, capacidade monetária, estudos e empregabilidade, de maneira a auxiliar a tomada de decisão no âmbito do retalho. Para tal serão utilizadas técnicas de prospeção de dados, em específico um método de particionamento de dados de acordo com a similaridade dos atributos anteriormente mencionados. Após a análise exploratória dos resultados da segmentação, serão aplicadas técnicas de analise espacial, recorrendo a um subconjunto de dados demonstrativos de consumo por clientes de maneira a determinar a fiabilidade das designações propostas para cada *cluster* (grupo).

2 – Estado de arte

2.1. Segmentação de mercado

Segmentação do mercado foi descrita por *Wendel* (1956), como, em certa medida, uma força que não pode ser ignorada, podendo resultar de tentativa e erro no sentido que programas generalizados de diferenciação de produtos podem se demonstrar como eficazes em alguns segmentos e ineficazes em outros.

Kamura e *Wedel* (2000) definem segmentação de mercado como um a divisão de um mercado de procura heterogénea, em sub-mercados com procura homogénea, permitindo a uma entidade retalhista adaptar a sua marca, produto ou serviço às necessidades do consumidor, valorizando assim estratégias de *marketing* diferenciadas. Semelhantemente, *Weinstein* (2004) caracteriza segmentação de mercado como o processo de particionar o mercado em grupos de potenciais consumidores, com características, tendências e necessidades similares, os quais irão, provavelmente, demonstrar hábitos de consumo semelhantes.

Segmentos de mercado podem ser caracterizados de diferentes de maneiras, de modo a explorar as preferências do público-alvo; preferências homogéneas, referindo-se a clientes com aproximadamente as mesmas preferências. Por outro lado, há preferências difusas no sentido que clientes variam na sua preferência e, finalmente, preferências agrupadas, que simbolizam segmentos de mercados que emergem de grupos de consumidores com preferências partilhadas (*Keller* e *Kotler*, 2009).

As segmentações de mercado são frequentemente compostas por variáveis que servem como base para o desenvolvimento de segmentações de mercados de consumo. *Kotler* (1998) distingue estas variáveis como: geográficas, demográficas, comportamentais, psicográficas, e aspetos relacionados com o produto.

Em suma, segmentação de mercado pode ser mencionada como um dos elementos chaves no marketing moderno e é o processo de dividir o mercado em diversos grupos e / ou segmento(s) baseados em fatores como demografia, geografia e fatores comportamentais e psicológicos. Assim os *marketeers* terão uma melhor compreensão do seu público-alvo, aumentando assim a eficácia do marketing (*Gunter* e *Furnham*, 1992)

2.2. Geodemografia

Segundo *Sleight* (1995) geodemografia é a análise de informação demográfica, que pode ser derivada dos dados dos censos da população ou de inquéritos em grande escala por unidade geográfica.

Brown (1991), refere-se a geodemografia como rótulos utilizados para o desenvolvimento de aplicações de segmentação (tipologias de área) que provam ser poderosos discriminantes de comportamentos de consumidor e servem de auxílio a analises de mercado.

Estes rótulos são a base da segmentação, descrevendo características de indivíduos ou grupos, sendo muitas vezes utilizados por *marketeers* para dividir o mercado em segmentos (*Sun*, 2009). Essas características podem ser resumidas a:

- Sociodemográficas – divisão dos consumidores em segmentos, que tem como base variáveis demográficas como a idade, género, rendimento, ocupação, classe social, educação e nacionalidade (*Armstrong* e *Kotler*, 2005);
- Geográficas – divisão dos consumidores em diferentes áreas geográficas, como ruas, cidades, estados, entre muitas outras;
- Estilo de vida – divisão dos consumidores de acordo com o seu estilo de vida, interesses e tendências (*Pickton* e *Broderick*, 2005).

2.3. Clustering

Análise de *clusters*, ou simplesmente *clustering*, é o processo de particionar um conjunto de objetos (ou observações) em subconjuntos. Também designado como segmentação de dados em algumas aplicações, o *clustering* particiona grande conjuntos de dados de acordo com a sua similaridade e é utilizado em diversas aplicações como inteligência de negócio (*business intelligence)*, reconhecimento de padrões em imagem, procura na web, biologia e segurança. Dentro de *business intelligence, clustering* pode ser utilizado para organizar grandes quantidade de clientes em grupos, onde os clientes dentro de um grupo partilham características similares. Estes grupos são denominados como *clusters* (*Han*, e *Kamber*, 2014)

No âmbito de aprendizagem automática (*machine learning*) *clustering* é considerado aprendizagem não supervisionada, ou seja, não existem segmentos predefinidos (*Tou* e *Gonzalez*, 1974), sendo assim uma forma de aprendizagem por observação, contrariamente a aprendizagem supervisionada onde a forma de aprendizagem é baseada em exemplos.

Quando considerada uma ferramenta de prospeção de dados (*data mining)* diferentes metodologias estatísticas são adaptadas como algoritmos, que podem ser diferenciados pela análise dos seus aspetos ortogonais. Estes aspetos são distinguidos e definidos por *Han* e *Kamber* (2000) da seguinte forma:

- Critérios de partição: em alguns métodos, todos os objetos são particionados de maneira a que não exista hierarquia entre *clusters*, isto é, todos os clusters estão no mesmo nível conceptual. Alternativamente, outros métodos particionam os objetos hierarquicamente, onde *clusters* podem ser formados em diferentes níveis de semântica.
- Separação de *clusters:* alguns métodos particionam os objetos em *clusters* mutualmente exclusivos. Em outras situações os *clusters* podem não ser exclusivos, isto é, um objeto pode pertencer a mais que um *cluster*.
- Medida de similaridade: diversos métodos determinam a similaridade entre dois objetos como a distancia entre eles. Tal distancia pode ser definida num espaço euclidiano, uma rede de estradas, um vetor no espaço, ou qualquer outro espaço. Em outros métodos, a similaridade pode ser definida como a conectividade baseada em densidades ou contiguidade.
- Espaço de *clustering*: muitos métodos de *clustering* procuram *clusters* dentro de todo o espaço existente nos dados. Estes métodos são uteis para conjuntos de dados com baixa dimensionalidade (ou seja, tem um baixo número de dimensões). Os restantes procuram *clusters* dentro de diferentes subespaços dentro dos mesmos dados, descobrindo *clusters* e subespaços (muitas vezes de com baixa dimensionalidade) que manifestem similaridade de objetos.

De acordo com Salvador e *Chan* (2003) existem quatro categorias de algoritmos de *clustering.* Essas categorias são descritas por *Han* e *Kamber* (2000) como:

- Particionais: os dados são divididos em *k* grupos de maneira a que exista pelo menos um objeto por *cluster*, estes algoritmos normalmente são baseados numa medida de similaridade.
- Hierárquicos: decompõem os dados hierarquicamente (de acordo com uma medida de distância), estes algoritmos podem ser aglomerativos ou divisivos:
    - Aglomerativo – sucessivamente junta objetos ou grupos que estejam perto uns dos outros até todos os objetos estarem agrupados ou ser cumprida a condição de paragem;
    - Divisivo – começa com os objetos todos no mesmo *cluster*, sendo depois progressivamente divido em *clusters* mais pequenos, até que cada objeto pertença a um *cluster* ou a condição de paragem seja cumprida.

- Baseados em densidade: um dado *cluster* continua a crescer desde que a densidade (número de objetos) em determinada vizinhança exceda um determinado limite.
- Baseado em grelha: efetuam uma quantização o espaço dos objetos num número finito de células que formam uma estrutura de grelha, sendo os *clusters* formados numa estrutura de grelha.

Estas técnicas são muito utilizadas em diferentes contextos. Dentro da segmentação de mercado existe uma panóplia de pesquisas que utilizam estas técnicas para dividir diversos conjuntos de dados em segmentos tipológicos.

Por exemplo *Wymer, et al* (1985) utilizam o *Super-CCP Algorithm for Large-Scale Area-Based Classifications* para particionar 130000 distritos enumerados com informação da geodemografia da população e criar 13 diferentes *clusters*.

Debenham, (2002) utiliza maioritariamente dados da geodemografia dos censos de 1991 de *Yorkshire* e o *Humber*, à escala do código postal (*postal code, postcode)*, para dividir a população em nove *clusters* com recurso ao algoritmo *k-means.*

Reeds e Vickers (2007) utilizam a geodemografia, proveniente dos censos do Reino Unido, e o algoritmo *k-means* para distribuir 223060 unidades territoriais em uma hierarquia de sete, 21 e 52 *clusters* (em diferentes escalas).

Algumas referências de interesse no âmbito da segmentação de mercado são a *Experian Mosaic*, a *CACI Acorn* e *ESRI Tapestry*, que serão superficialmente analisadas com base na metodologia disponibilizada no *website* de cada entidade.

A *Experian Mosaic* (*Experian*, 2014) é uma classificação de consumidores projetada para ajuda a compreender a demografia, estilo de vida, preferências e comportamentos da população adulta do Reino Unido em um detalhe extraordinário. Esta classificação é baseada em mais de 850 milhões fontes de dados, na qu al 28% são pertencentes aos censos, com mais de 450 variáveis. Diversas técnicas estatísticas são utilizadas durante o processamento da informação, como por exemplo a *Singular Value Decomposition, Hitwise e Cheetahmail*. O *Mosaic* divide o Reino Unido em 238 subtipos, divididos de 66 tipos, permitindo um elevado nível de descriminação de indivíduos.

A *CACI Acorn* (CACI, 2018) descreve-se como uma poderosa ferramenta de segmentação que combina geografia com demografia e informação sobre estilos de vida e a localização onde as pessoas vivem, suas características e comportamento subjacentes, criando uma ferramenta para compreender diferentes tipos de pessoas em diferentes áreas por todo o Reino Unido. Isto permite aos utilizadores compreender o tipo de pessoas que vivem em cada área, aumentar mercados ou a quantidade de serviços disponíveis. Os dados utilizados, privados e públicos, incluem, por exemplo, o registo de terras, censos, fontes comerciais de informação sobre as idades dos residentes, dados de benefícios, densidade populacional, casas de assistência, habitação social, entre outros. A *Acorn* particiona códigos postais em 6 categorias, 18 grupos e 62 tipologias, com as tipologias dividas em 313 micro-segmentos. A abordagem da *CACI* à geodemografia começa com a separação do processo de definição das tipologias usadas para descrever a nossa sociedade do processo de atribuir códigos postais a essas tipologias. Esta abordagem permite utilizar diferentes algoritmos durante o processo de atribuição. A vantagem principal é que, mesmo que não exista melhor alternativa, será sempre possível atribuir uma tipologia *Acorn* utilizando a abordagem tradicional, garantindo uma melhoria em geral.

A *ESRI Tapestry* (ESRI, 2014) é um sistema de segmentação geodemografica que integra atributos de consumidores com características residenciais para identificar mercados e classificar vizinhanças (*neighborhoods)* em 67 segmentos de mercado comportamentais distintos. Para tal é utilizada uma combinação de técnicas, como o algoritmo de partição iterativa *k-means*, para criar os *clusters* iniciais ou segmentos de mercado, seguido da aplicação do método hierárquico da mínima variância de *Ward* para agrupar os *clusters*. Os dados utilizados incluem os censos de

2010 dos Estados Unidos, a *American Community Survey (ACS)*, as atualizações de dados demográficos da *ESRI*, a *Experian's ConsumerView Database,* entre outras fontes como inquéritos e pesquisas, sendo estas posteriormente expostas a diversas metodologias estatísticas multivariadas.

2.4. Normalização dos dados

Todas as técnicas de *clustering* são baseadas na similaridade ou dissimilaridade dos objetos que se quer agrupar. Esta é medida pela construção de uma matriz de distâncias, que reportam a distância entre pares de casos (unidades territoriais) para cada variável. É claro que problemas podem surgir se existirem diferentes escalas ou magnitudes entre as variáveis. Em geral, variáveis com maiores valores e maior variação vão ter mais efeito na medida final de similaridade. (*Vickers* e *Rees*, 2007). Esta será uma operação necessária devido às diferenças na magnitude de valores apresentados no conjunto de dados utilizado e uma consequência da utilização de distâncias euclidianas como métrica de similaridade.

Vaishali e *Rupa* (2011) referem que técnicas de pré processamento *são aplicadas a dados não tratados para tornar os dados limpos, livres de ruido e consistentes. Normalização de dados estandardiza os dados não tratados convertendo-os em intervalos específicos através de transformações lineares que podem gerar clusters de boa qualidade e melhorar a precisão de algoritmos de clustering.* Não existe uma maneira universal para normalizar dados e assim a escolha é deixada em grande parte ao critério do utilizador (*Karthikeyani* e *Thangavel*, 2009).

Inicialmente foram consideradas três métodos de normalização: *z-score, min-max* e transformação logarítmica. A transformação logarítmica verificou-se não exequível devido à presença de valores zero no conjunto de dados utilizado. *Mohamad* e *Usman* (2013) concluem na sua pesquisa que, dos métodos propostos, onde está incluída a metodologia *min-max*, que o método que obtém resultados mais precisos e eficientes, com a utilização do algoritmo *k-means*, é o *z-score*.

O *z-score* é uma técnica de normalização, utilizada neste projeto, que transforma variáveis com distribuição normal numa distribuição normal padronizada. Dado um conjunto de dados Y, a fórmula de normalização do *z-score* é definida como:

$$x_{ij} = Z(x_{ij}) = \frac{x_{ij} - \bar{x}_j}{\sigma_j} \qquad (1)$$

onde, $\bar{x}_j$ e σ_j são a média e desvio padrão da j-ésima variável, respetivamente. As variáveis transformadas têm uma média de 0 e variância de 1.

É necessário aplicar uma normalização aos dados utilizados, pois de outra forma os intervalos de valores de cada variável iriam agir como ponderações (pelos motivos descritos anteriormente), o que iria prejudicar o processo de agrupamento dos dados.

3 – Metodologia

3.1. Fontes de informação

Para a realização deste projeto foram utilizados maioritariamente dados públicos, provenientes dos censos da população fornecidos pelo INE (Instituto Nacional de Estatística). Esta informação gratuita encontra-se disponível em https://ine.pt e encontra-se exposta na tabela 1.

Tabela 1 – Informação utilizada e sua caracterização.

Designação da fonte	Designação	Formato	Descrição
Base Geográfica de Referenciação de Informação	BGRI	CSV	Dados relativos à população de Portugal Continental
Carta Administrativa Oficial de Portugal	IGEO	SHP	Carta administrativa de Portugal
Dados Nacionais Estatísticos	INE	CSV	Dados relativos à população de Portugal
Base de dados Portugal Contemporâneo	PORDATA	XLS	Dados relativos à população de Portugal
Base de dados da ESRI	ESRI	SHP	Dados relativos a rendimentos da população de Portugal

Os censos são uma fonte de dados importante quando se pretende segmentar uma população visto que fornece características de áreas locais com uma escala geográfica adequada ao tipo de estudo que se pretende, reduzindo os efeitos de agregação e consequente perda de informação. Estes possuem dados contextuais que demonstram a com precisão as principais características demográficas da população, fornecendo informação valiosa como vizinhanças, diversidade étnica, emprego, indústria, níveis de educação, saúde, entre outros.

Os dados referentes aos censos utilizados, são maioritariamente da Base Geográfica de Referenciação de Informação (BGRI), disponibilizada com os censos de 2011. A BGRI é disponibilizada pelo Instituto Nacional de Estatística (INE), em formato *shapefile*, e corresponde a uma subdivisão do território nacional ao nível da subsecção estatística em que cada subsecção é descrita por 122 atributos (denominadas como variáveis neste relatório).

Finalmente, foi utilizado o serviço *Enrich Layer* (https://doc.arcgis.com/en/arcgis-online/analyze/enrich-layer.htm), fornecido pela ESRI, para obter informação relativa aos rendimentos da população portuguesa, mais concretamente os quintis de rendimentos. Esta encontra-se descrita no subcapítulo 3.5.8 do relatório.

3.2. Software utilizado

Este projeto foi desenvolvido em ambiente *Windows 10* e recorreu ao seguinte software:

Tabela 2 - Software utilizado

Designação	Desenvolvedor	Versão	Website
ArcGIS Pro	ESRI	2.1	https://pro.arcgis.com/
Python	Python	3.6	https://www.python.org/
Pandas	NumFOCUS	0.23.3	https://pandas.pydata.org/
Scikit-learn	INRIA	0.19.1	scikit-learn.org/

O ArcGIS *Pro* é um sistema de informação geográfica, distribuído pela *ESRI*, que permitiu desenvolver aspetos deste projeto como a análise espacial efetuada, a segmentação proposta, a visualização e processamento da informação. O *python* é uma linguagem de

programação, e foi utilizada para processar grande parte da informação inicialmente recolhida, realizar testes e determinar as variáveis utilizadas. O *pandas* e *scikit-learn*, ambas bibliotecas de *Python*, permitiram o manuseamento de informação em formato tabular e a utilização de diferentes funções e algoritmos, respetivamente.

3.3. K-*means*

De entre os diversos métodos de segmentação existentes, os métodos *post-hoc*, especialmente os de *clustering*, são ferramentas de análise poderosas e frequentemente usadas na prática. (Dillon et al., 1993; *Wedel* e *Kamakura*, 1998). Dentro dos processos de *data mining*, o *clustering* é um dos métodos mais eficazes de identificar distribuições de interesse e padrões diferentes dentro de um conjunto de dados.

O *clustering* assume várias designações de acordo com o contexto onde estão inseridos, tal como aprendizagem não supervisionada (reconhecimento de padrões), taxonomia numérica (biologia, ecologia), tipologia (ciências sociais) e partição (teoria de grafos) (*Theodoridis* e *Koutroubas*, 2008).

Existem vários métodos de *clustering* disponíveis, variando em aspetos como os parâmetros de entrada, escala, casos de uso e geometria. É importante notar que diferentes métodos vão produzir diferentes resultados, consoante os dados, tornando difícil a escolha do melhor método a utilizar. Neste relatório escolheu-se o método mais frequentemente usado em estudos de segmentação de mercado, o *k-means clustering* (*A. K. Jain*, 2010).

O método *k-means* é uma das técnicas mais simples para a criação de agrupamentos que otimizem a função de critério de qualificação, definida globalmente ou localmente (*Vaishali e Rupa*, 2011), sendo um dos métodos mais utilizados em indústrias que utilizam a geodemografia (*Harris et al.*, 2005).

O algoritmo *k-means* encontra uma partição de forma a que a soma dos quadrados dos erros produzidos pela diferença entre a média empírica de um *cluster* e o valor médio dos atributos dos objetos pertencentes a esse mesmo *cluster* seja minimizada. Considerando a equação, ***x*** representa um conjunto com ***i***, …, ***n*** objetos com ***d***-dimensões que serão agrupados em ***k*** *clusters*, pertencentes a um conjunto ***C*** com 1, …, K *clusters*, o quadrado dos erros determinados pela diferença entre a média do *cluster* ***k***, μ_k, e os objetos do *cluster* C_k, é definida por (equação 2) (*A. K. Jain*, 2010)

$$SQE(C_k) = \sum_{xi \in C_k} ||x_i - \mu_k||^2 \qquad (2)$$

Desta forma, a função que se pretende minimizar é descrita por (equação 3) (*A. K. Jain*, 2010):

$$SQE(C) = \sum_{k=1}^{K} \sum_{xi \in C_k} ||x_i - \mu_k||^2 \qquad (3)$$

Para tal, o algoritmo efetua o processo descrito de seguida:

Input:

- *k:* número de *clusters*,
- *D:* conjunto de dados com ***n*** objetos.

Ouput: Conjunto de *clusters*

Método:

(1) Arbitrariamente escolher *k* objetos de *D* para determinar os centroides iniciais;

(2) Repetir
(3) (re)atribuir cada objeto ao *cluster* que mais similar for ao objeto, baseado no valor médio dos objetos do *cluster*;
(4) (re)calcular a média de cada *cluster*, ou seja, calcular a média dos objetos para cada cluster;
(5) Termina quando, de uma iteração para outra, não há alterações na atribuição do objeto a um *cluster*.

Este processo é ilustrado na Figura 1 (*Piech*, 2013), de seguida apresentada:

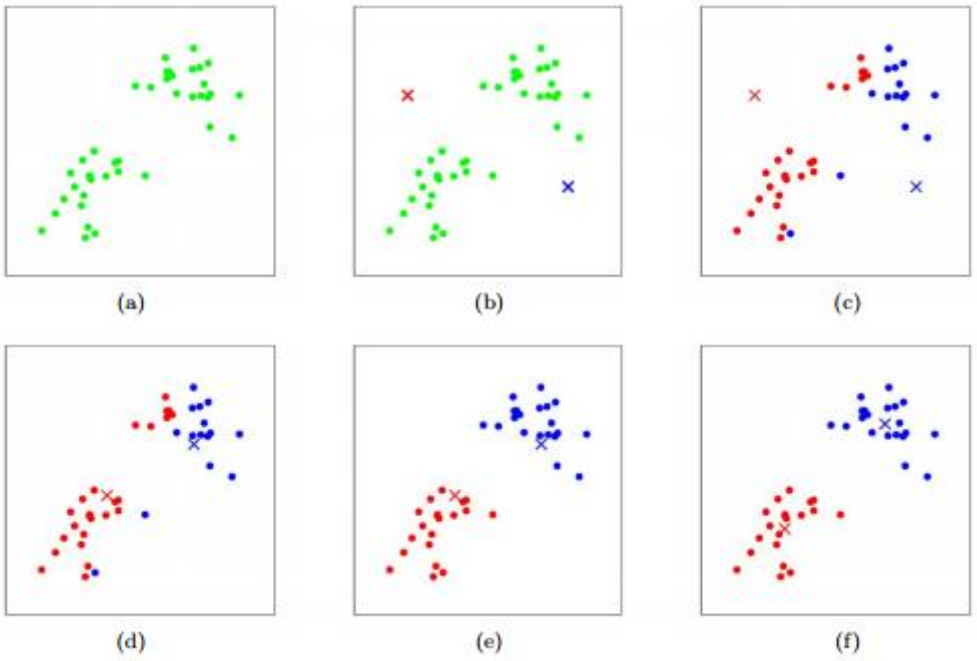

Figura 1- Algoritmo k-means. Exemplos de treino são mostrados como círculos e os centroides dos clusters como cruzes. (a) Dados originais. (b) Centroides aleatórios dos clusters iniciais. (c-f) Ilustração de duas iterações do k-means. (Piech, 2013),

Neste exemplo da aplicação do algoritmo *k-means* (Figura 1), inicialmente os objetos encontram-se não particionados (a), são escolhidos arbitrariamente os centroides dos *clusters* iniciais (b) e cada objeto é atribuído ao centroide mais próximo (c). Finalizado este processo é recalculado o valor dos centroides (d) e verifica-se se existem alterações na atribuição de objetos ao centroide mais próximo (e). Caso não exista alteração na atribuição de objetos o algoritmo termina (f), caso exista o algoritmo volta a calcular a posição do centroide e o processo é repetido até que não exista alteração na atribuição de objetos ou chegue a número máximo de iterações. De notar que a escolha arbitraria dos centroides (b) é feita pela escolha de uma observação e não por um posicionamento aleatório, sendo uma gralha da figura original.

O algoritmo utilizado requer três parâmetros de entrada definidos pelo utilizador: o número de *clusters* (*k*), a inicialização de *clusters* e a métrica de distâncias. O número de *clusters* indica ao algoritmo em quantas partições o utilizador pretende dividir a informação. A inicialização de *clusters* é uma opção que permite especificar as localizações iniciais dos *clusters*, sendo uma opção utilizada, por exemplo, quando existe um conhecimento prévio das tendências da informação em análise, potenciando uma melhoria nos resultados. Por fim, a métrica de distâncias é o espaço métrico que permite determinar a similaridade entre pares de objetos, baseada no quão perto ou distantes estes se encontram entre si. Neste trabalho a operação de *clustering* foi feita recorrendo à ferramenta *Multivariate Clustering*, do *ArcGIS Pro (*http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/multivariate-clustering.htm).

A escolha de *k*, a única obrigatória das três, será explorada no subcapítulo 3.5 mais detalhadamente, devido a sua importância. A inicialização de *clusters* pode produzir resultados finais diferentes, visto o algoritmo convergir a um mínimo local (não garante o mínimo global). A ferramenta utilizada permite a opção *Optimized seed locations*, escolhendo aleatoriamente a primeira *seed* (centroide) e garantindo que as *seeds* subsequentemente selecionadas representam objetos que se encontram distantes entre si no espaço de dados (valores dos atributos) garantindo a captura de diferentes áreas, melhorando assim a performance do algoritmo.

Por fim, é necessário definir uma métrica de distância de maneira a quantificar a similaridade entre objetos, visto esta influenciar a maneira e ordem como o algoritmo irá agrupar os objetos. Existem diversas métricas de distância como a euclidiana, Mahalanobis, Itakura-Saito, entre outras. A métrica de distância mais comumente utilizada em conjunto com o algoritmo *k-means* é a euclidiana (*Liu et al.*, 2012, *A. K. Jain*, 2010), tendo sido a escolhida para este projeto.

3.4. Seleção de variáveis

Um dos problemas da seleção de variáveis para um projeto de *clustering* de geodemografia é a escassez de estudos que comprovem quais as melhores variáveis para estudos de geodemografia (*Webber, R.*, 2004).

Dentro da análise de *clusters*, citando *Kaufman* e *Rousseauw* (1990) *deve ser notado que uma variável que não contenha qualquer informação relevante é pior que inútil, porque fará o clustering menos aparente. A ocorrência de várias "variáveis lixo" ira matar o processo de clustering porque irá introduzir muitos termos aleatórios em distâncias, ocultando assim informações uteis provenientes de outras variáveis.*

As variáveis consideradas foram selecionadas de acordo com os objetivos do projeto, disponibilidade e capacidade de adicionar definição aos diferentes *clusters*, utilizando outros estudos e sistemas geodemograficos comerciais atuais como referência (i.e *Experian Mosaic*, *CACI Acorn).* O objetivo da seleção de variáveis é escolher o menor número possível de variáveis que satisfatoriamente representem as diferentes dimensões consideradas (*Bailey, et al.,* 2000).

No âmbito deste projeto, foram inicialmente recolhidas 134 variáveis (Anexo 3), na forma de contagem (por exemplo o número de indivíduos por género). Estas, por sua vez, foram expostas a um processo de seleção de duas fases, descrito de seguida:

I. Cada variável candidata foi relacionada a uma contagem base correspondente. Caso várias variáveis candidatas partilhem a contagem base é escolhida a que mais expressividade atribuir a cada *cluster*. O intuito deste processo é reduzir o número de variáveis muito correlacionadas e limitar a redundância nos dados.
II. Foram escolhidas diferentes combinações de variáveis candidatas (não mutualmente exclusivas) e criados diferentes conjuntos de *clusters*. Posteriormente foi calculado um índice de validade para cada conjunto sendo escolhida a combinação que apresentou o maior índice, mantendo variáveis indispensáveis devido a sua contribuição explicativa para o retalho (por exemplo distribuições de idades). O índice de validade utilizado foi o *silhouette score*, disponibilizado pela biblioteca *Scikit-learn* (http://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html).

No âmbito de *clustering* é possível distinguir dois tipos de índices de validade para medir a qualidade dos resultados obtidos: índices externos e índices internos. Um índice de validade externo é uma medida de concordância entre duas partições em que a primeira partição é a estrutura de *clustering*, conhecida *a priori*, e a segunda a resultante do procedimento de *clustering*. (*Dudoit et al.*, 2002). Índices de validade internos são usados para medir a qualidade de uma estrutura resultante do procedimento de *clustering*, sem a existência de informações externas (*Tseng el al*,.2005).

Visto não existir um conhecimento *a priori* da estrutura esperada, foi utilizado um índice de validade interno. Em geral, os índices internos avaliam a separação dos diferentes *clusters* (extra-cluster) e a compactação de cada *cluster* (intra-*cluster*). Como exemplos destes índices temos o índice de *Dunn*, o índice de *Davies–Bouldin* e o índice da silhueta (*silhouette score / índex*), tendo este último sido escolhido como métrica de qualidade neste projeto.

Segundo *Lletí, el al* (2004) *as silhuetas de clusters são suficientemente sensíveis para determinar a presença de variáveis redundantes num conjunto de dados*. O valor de silhueta determina o quão similar um objeto é com os outros objetos no seu *cluster* (coesão) quando comparado com objetos em outros *clusters*, de acordo com uma métrica de distância escolhida e é calculado pela equação 4 (*LLetí, et al*, 2004):

$$s(i) = \frac{b(i) - w(i)}{\max\{\mathrm{b}(i), w(i)\}} \qquad (4)$$

Com:

$$b(i) = \min\{B(i,k)\} \qquad (5)$$

Onde w(i) é a distância média do i-ésimo objeto a objetos no mesmo *cluster*, b(*i*) a distância média mínima do i-ésimo objeto a todos os objetos de qualquer *cluster* que não incluam *i*, e B (i, k) a distância média do i-ésimo objeto a objetos em diferentes *clusters*. A média de todos os pontos de um *cluster* ilustra o quanto coeso um *cluster* é, logo a média de um conjunto de dados é uma medida do quão apropriadamente os dados foram agrupados. Esse valor pode ser estimado da seguinte forma (Equação 6) (*LLetí, et al*, 2004):

$$\bar{s}(k) = \frac{\sum_{i=1}^{m} s(i)}{n} \qquad (6)$$

Onde *n* denota o número de objetos num conjunto de dados.

Os valores do *silhouette índex* estão compreendidos entre 1, o que indica que os objetos estão muito distantes dos *clusters* vizinhos, e -1, assinalando que os objetos foram provavelmente atribuídos ao *cluster* errado, com o valor 0 indicando que existe ambiguidade na atribuição de objetos a *clusters*. *Kaufmann* e *Rousseuw* (1990) apresentam uma interpretação subjetiva do *silhouette índex*, sendo esta independente do número de objetos, como se apresenta listada na tabela 3:

Tabela 3 - Valores de silhueta e interpretação

Silhouette index	Interpretação proposta
≤ 0.25	Não foi encontrada uma estrutura substancial
0.26 - 0.50	A estrutura é fraca e pode ser artificial; utilizar outros métodos neste conjunto de dados
0.51 - 0.70	Uma estrutura razoável foi encontrada
0.71 - 1.00	Uma estrutura forte foi encontrada

A análise do *silhouette índex* no âmbito da operação de *clustering* realizada neste trabalho, permitiu selecionar 48 variáveis com um *silhouette índex* de 0.54, utilizando a distância euclidiana como métrica de similaridade. No subcapítulo 3.6 é feita uma análise das variáveis escolhidas, e uma descrição das mesmas pode ser encontrada no anexo 1. As variáveis foram agrupadas em grupos taxonómicos, descritos de seguida:

I. Agregado familiar – conjuga informação sobre a dimensão de uma família e a sua condição perante a habitação;

II. Emprego – refere-se aos diferentes sectores que empregam a população e ocupação de não trabalhadores;
III. Divisões dos alojamentos – número de divisões por alojamento;
IV. Dimensões dos alojamentos – dimensão dos alojamentos;
V. Educação – nível mais alto de educação completado por um individuo;
VI. Época de construção – indica o ano no qual a habitação foi construída;
VII. Sexo por idade – idade de um indivíduo, de determinado gênero, durante uma data especifica;
VIII. Quintis de rendimentos – conjunto de rendimentos ordenados em cinco partes iguais;
IX. Densidade populacional – número de indivíduos por unidade de área.

3.5. Número de *clusters* (*k*)

Um dos parâmetros do algoritmo *k-means* é o número de *clusters* desejado pelo utilizador, normalmente designado por *k*. Este parâmetro influencia a maneira como o algoritmo particiona a informação, o que irá provocar diferenças no resultado obtido. Apesar da sua importância, não existe uma metodologia matemática perfeita para a sua determinação, recorrendo-se normalmente a diferentes heurísticas de maneira a tentar estimar esse valor. Essas heurísticas podem por vezes ser contraditórias para a mesma aplicação, assim a escolha de *k* deve ser baseada na análise (humana) dos agrupamentos formados de maneira a perceber se representam os dados e o objetivo que o *clustering* pretende alcançar (*Vickers e P. Rees*, 2001).

Neste projeto foram consultadas duas heurísticas, o *silhouette score* (explorado no capítulo anterior) e o *elbow method (Thorndike R. L.*, 1953*)*. Para o desenvolvimento destas heurísticas foram utilizadas as varáveis descritas no subcapítulo 3.6. O *elbow method* é um método de interpretação e validação da consistência de *clusters*, baseado na soma dos erros de cada *cluster* em função do número de *clusters*. O número de *clusters* aconselhado por esta heurística é identificado por uma estabilização na queda do valor da soma destes erros, provocando um angulo visual, como podemos observar na Figura 2:

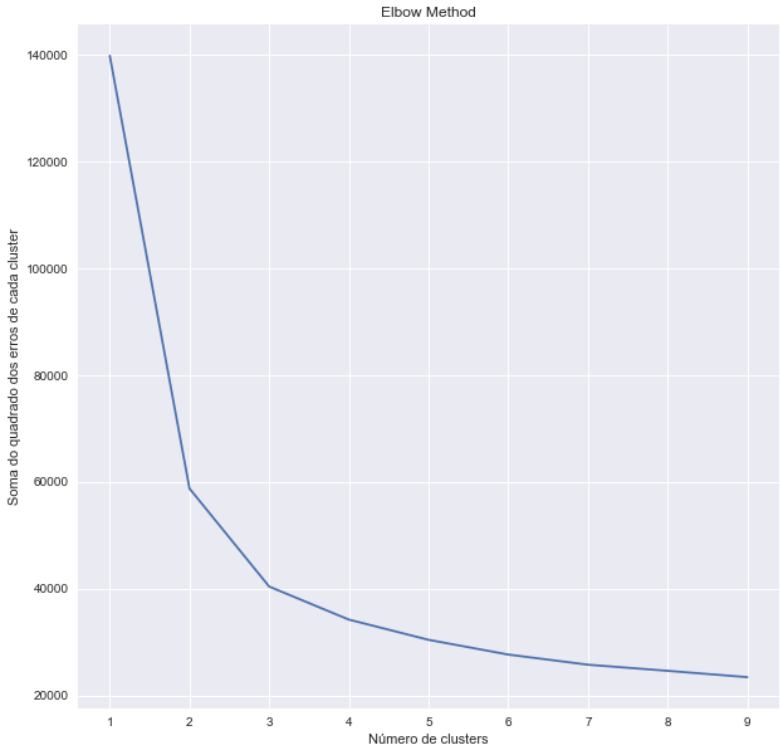

Figura 2 - Elbow method

Podemos verificar que o último ponto de inflexão verificado no gráfico da Figura 2 encontra-se quando o número de *clusters* é 3.

Para a determinação do número de *clusters* através do *silhouette* score, foram calculados os índices para $k = \{3, 4, 5, 6\}$, como pode se verificar na tabela 4:

Tabela 4 - Silhouette score para diferentes valores de k

Número de clusters	*Silhouette* Score
3	0.738
4	0.601
5	0.541
6	0.479

Podemos observar que o valor do índice vai diminuindo face à adição de *clusters.* Desta forma, podemos concluir que o número de *clusters* ótimo estimado por esta heurística é 3.

Apesar das heurísticas consultadas neste projeto indicarem um *k* ótimo de 3 *clusters*, quando efetuada uma análise exploratória para $k = \{4, 5\}$ foram observáveis diferenças, entre *clusters,* suficientes para ignorar o resultado de ambas as heurísticas, tendo sido escolhido um *k* de 5. A análise exploratória destas diferenças é feita no subcapítulo 3.6.

3.6. Análise e tipologia do *clustering* de freguesias

As 2882 freguesias de Portugal Continental foram agrupadas em 5 *clusters*, sendo então possível examinar as médias de cada *cluster*, por dimensão (variável) de maneira a avaliar as diferenças entre *clusters* (*Everitt et al.*,2001). Relativamente à sua distribuição (Tabela 5):

Tabela 5 - Distribuição de freguesias por cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Nº de freguesias	521	2089	65	167	37
Percentagem (%)	18.5	72.4	2.2	5.7	1.2

Podemos observar na tabela 5 que o *cluster* 2 é constituído por 72% das freguesias e os *clusters* 3 e 5 com apenas 2% e 1% respetivamente. A sua distribuição espacial pode ser verificada na Figura 3:

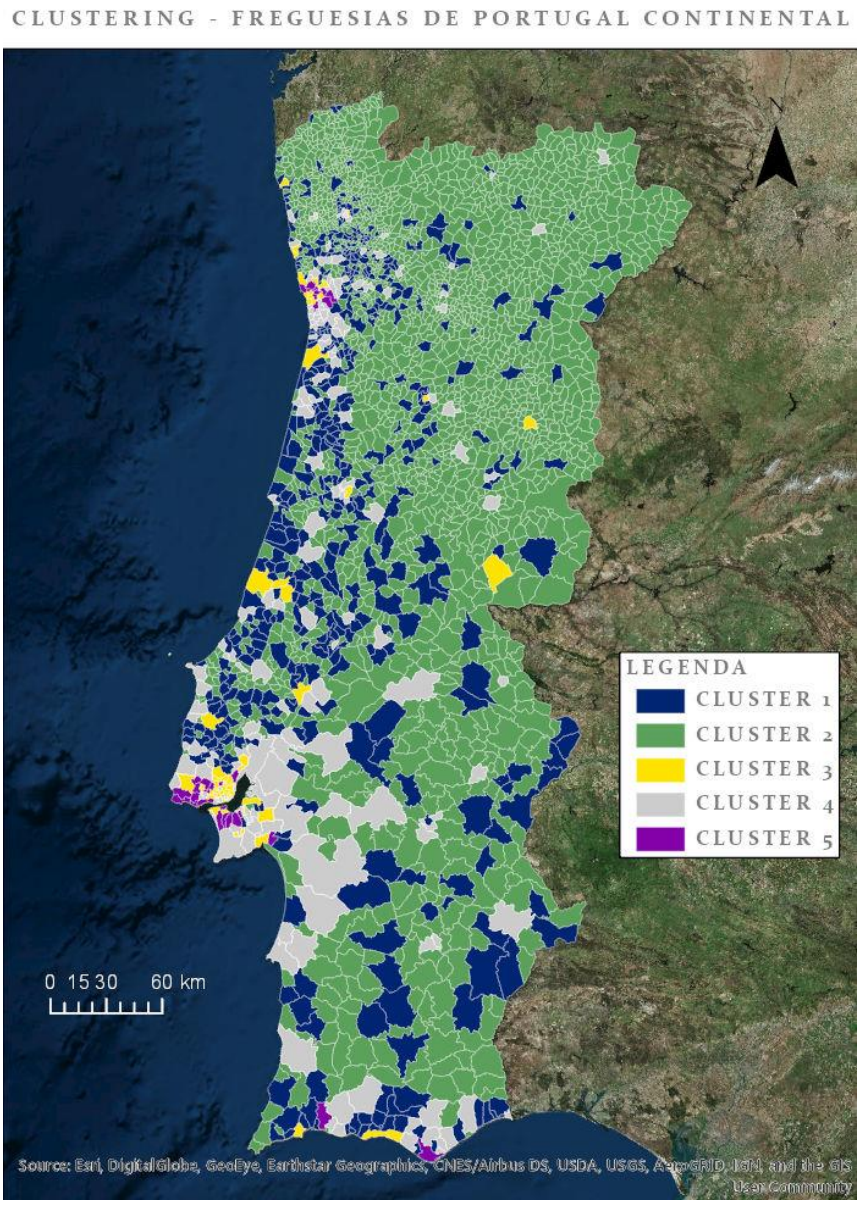

Figura 3 - Freguesias de Portugal Continental por cluster

De seguida será feita uma análise exploratória dos atributos utilizados para o desenvolvimento do *clustering*, de maneira a compreender quais as diferenças entre *clusters*.

3.6.1 - Agregado familiar

O agregado familiar descreve informação sobre a dimensão das famílias, a sua condição perante a habitação e presença de filhos.

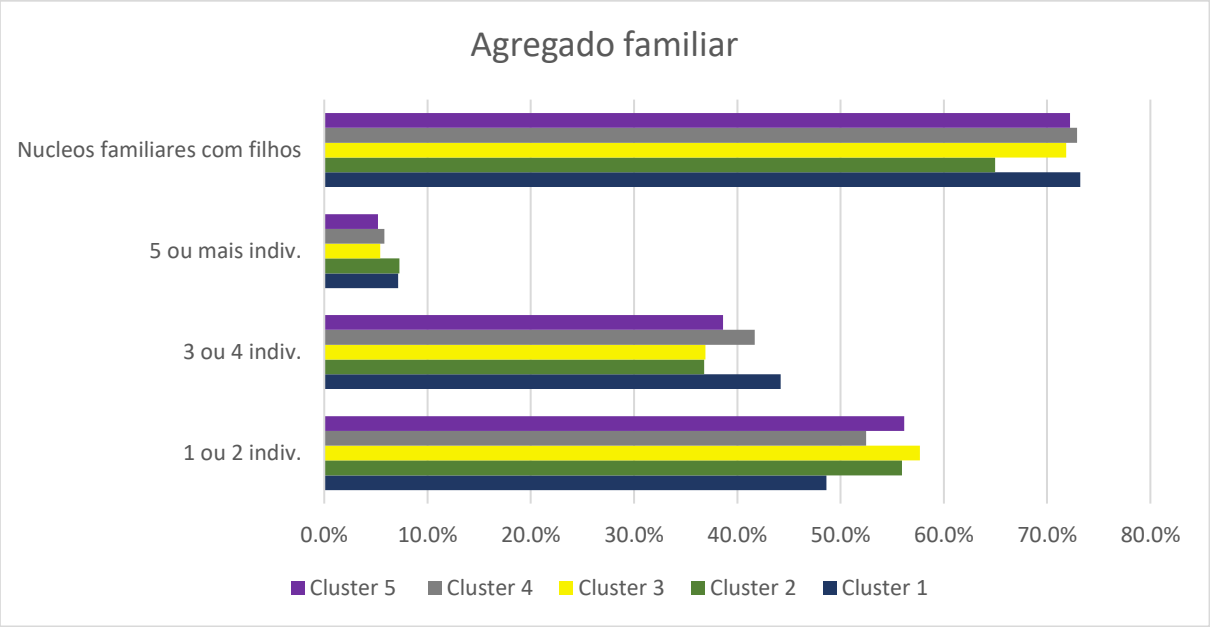

Figura 4 - Distribuição de variáveis sobre o agregado familiar

Em geral as famílias com 1 e 2 indivíduos dominam o tamanho do agregado familiar, com destaque para *cluster* 3 com a maior percentagem de famílias de pequena dimensão. Apesar do *cluster* 1 ser composto por freguesias onde a percentagem média para as famílias de 1 ou 2 duas pessoas é mais baixa que nos restantes *clusters*, este demonstra predominância para famílias com mais de 3 pessoas. Todos os agrupamentos demonstram uma quantidade considerável de núcleos familiares com descendência, com o *cluster* 2 substancialmente abaixo dos restantes.

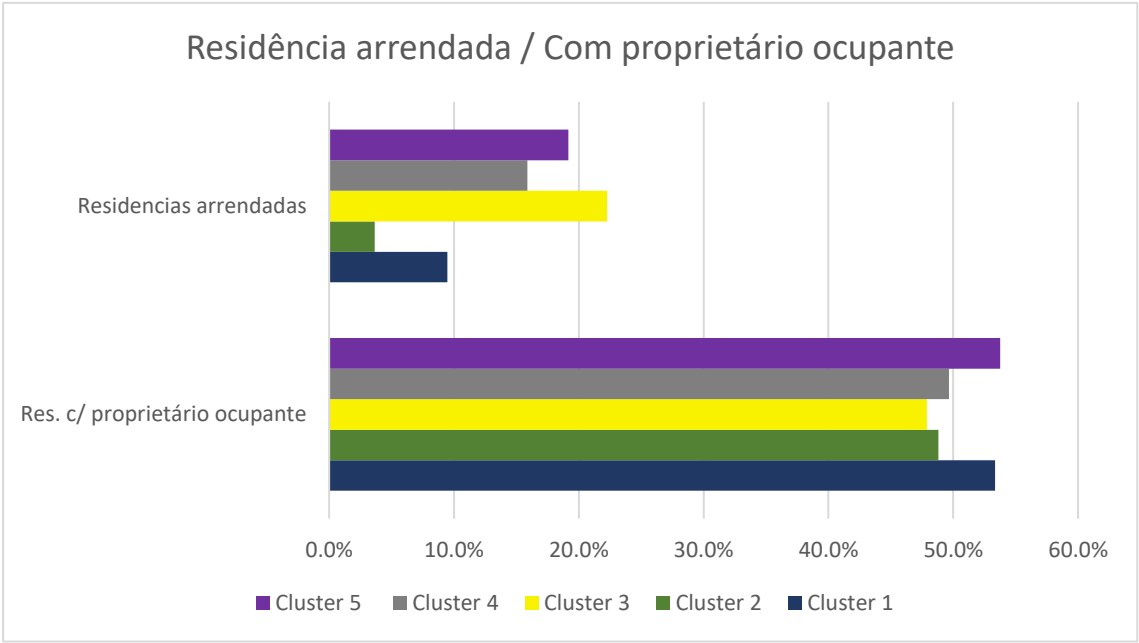

Figura 5 - Residência arrendada / Com proprietário ocupante

É possível observar pela Figura 5, que a maioria das residências são ocupadas pelo proprietário, com os *clusters* 1 e 5 a apresentarem predominância neste ponto. Relativamente a casas arrendadas, o *cluster* 3 apresenta a maior percentagem, com o *cluster* 5 com alguma relevância e o *cluster* 2 a apresentar valores substancialmente abaixo dos restantes.

3.6.2 - Emprego

As variáveis sobre o emprego especificam os sectores onde os indivíduos das freguesias trabalham e a ocupação de indivíduos não trabalhadores.

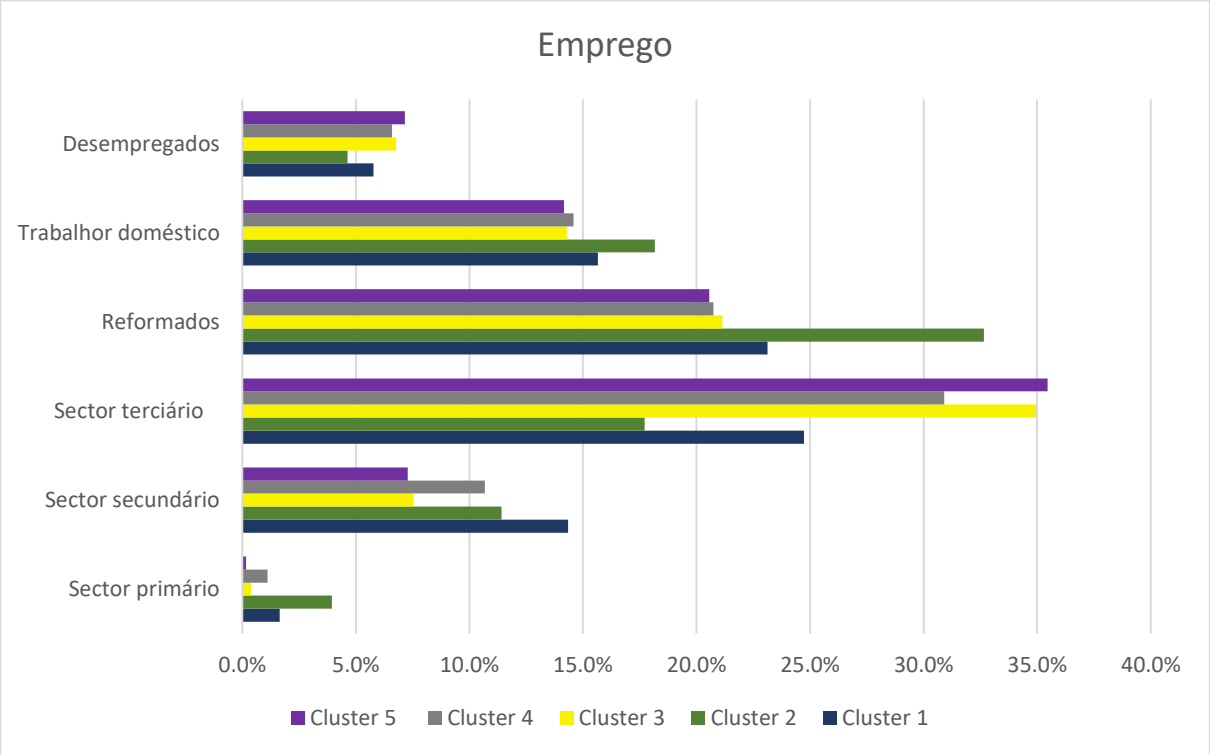

Figura 6 - Distribuição da população face ao emprego

Dentro das diferentes situações de empregabilidade dos indivíduos, é possível verificar que o sector terciário emprega a maioria dos habitantes, com os *clusters 5* e 3 com uma alta representatividade nesses setores e um número de desempregados acima da média. O *cluster* 2 apresenta predominância na percentagem de trabalhadores domésticos, reformados e empregados no setor primário. Por fim, o *cluster* 1 apresenta uma maior quantidade de pessoas empregadas no setor secundário, relativamente aos restantes agrupamentos.

3.6.3 - Divisões por alojamento

Estas variáveis têm o intuito de descrever a dimensão das habitações ocupadas pelos residentes das freguesias.

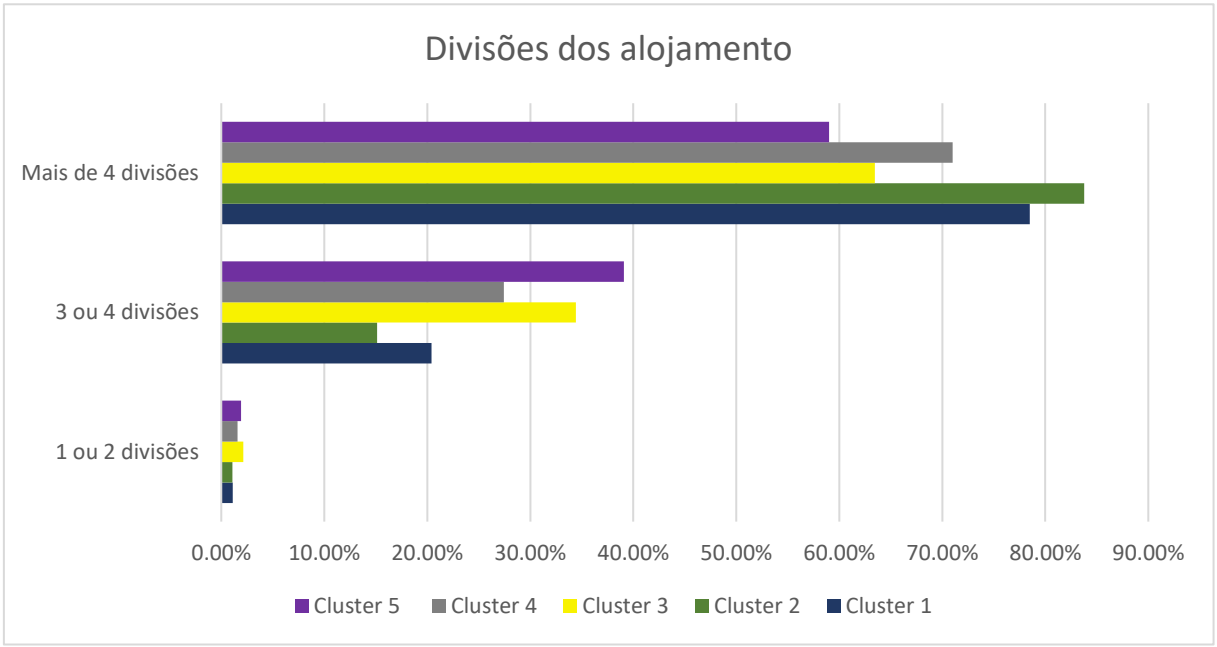

Figura 7 - Distribuição por diferentes quantidades de divisões por alojamento

É possível observar que a maioria das habitações têm mais de 4 divisões, apesar de se notar a existência de bastantes alojamentos com 3 ou 4 divisões nos *clusters* 3 e 5 assim como uma percentagem de habitações com 1 ou 2 divisões acima da média.

3.6.4 - Dimensão dos alojamentos

Semelhante às variáveis anteriormente exploradas, estas têm o intuito de aumentar a expressividade da dimensão dos alojamentos.

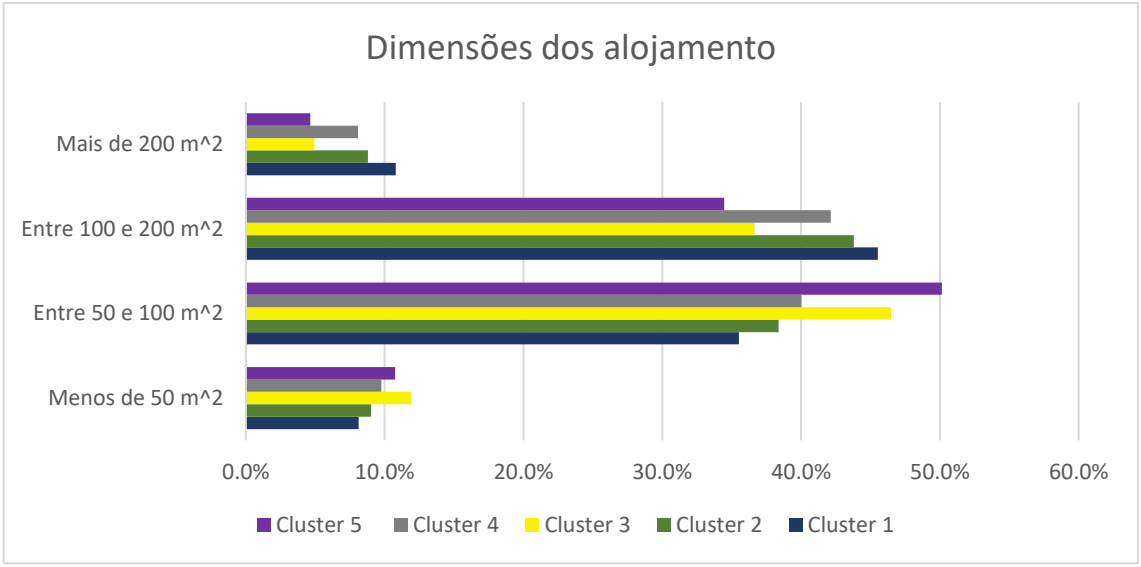

Figura 8 - Distribuição de diferentes dimensões de alojamentos

Em concordância com o número de divisões por alojamento, os *clusters* 3 e 5 apresentam a maior percentagem de pequenas habitações (menos de 50 m^2 e entre 50 e 100 m^2). Contrariamente, os *clusters* 1 e 2 contêm freguesias com uma elevada percentagem de grandes habitações (entre 100 e 200 m^2 e mais de 200 m^2). É possível ainda verificar que o *cluster* 4 apresenta sempre valores médios para os diferentes tamanhos de habitação.

3.6.5 - Educação

As variáveis sobre a educação permitem analisar a distribuição relativamente à alfabetização e completude de um ciclo de estudos.

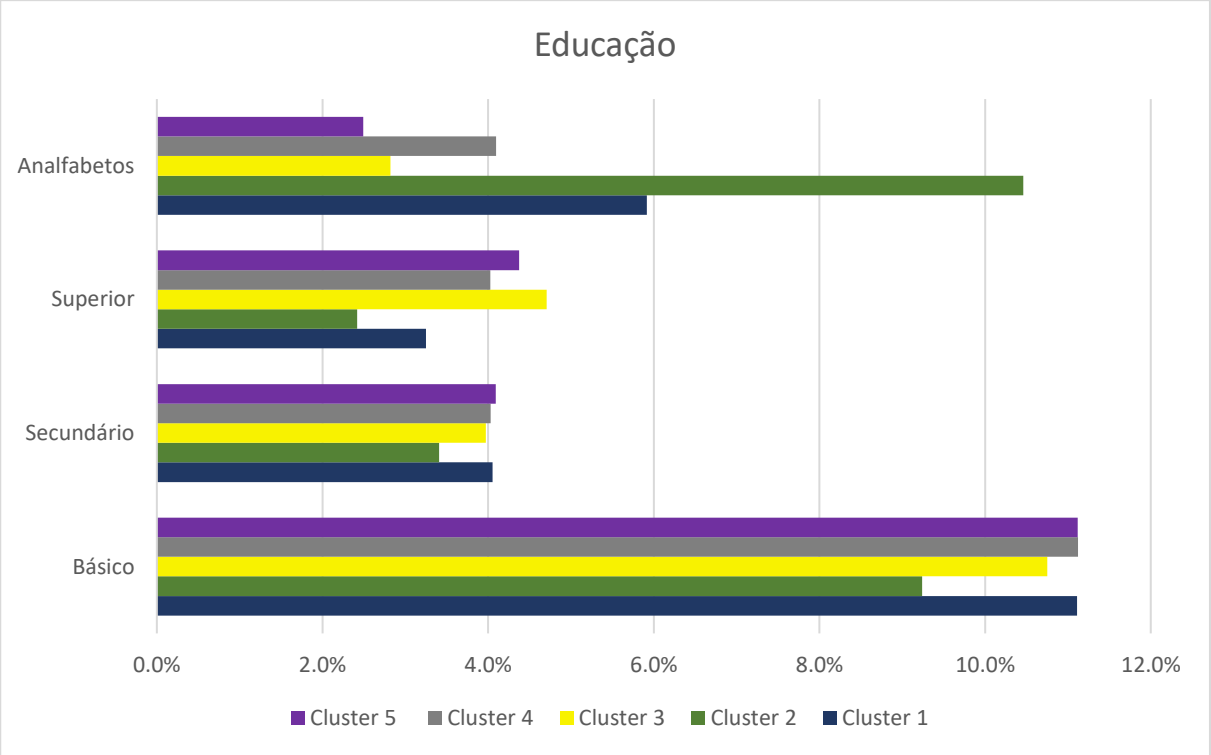

Figura 9 - Distribuição dos diferentes graus de habilitação académica

Para os diferentes níveis de educação, verifica-se um elevado número de analfabetos no *cluster* 2 e consequentemente a menor percentagem de níveis de educação em geral. Os *clusters* 1 e 5 demonstram o maior número de indivíduos com o secundário completo e o *cluster* 3 o maior número de pessoas com um curso superior completo.

3.6.6 - Época de construção

Este atributo permite saber o intervalo temporal no qual houve mais construção de imóveis habitacionais.

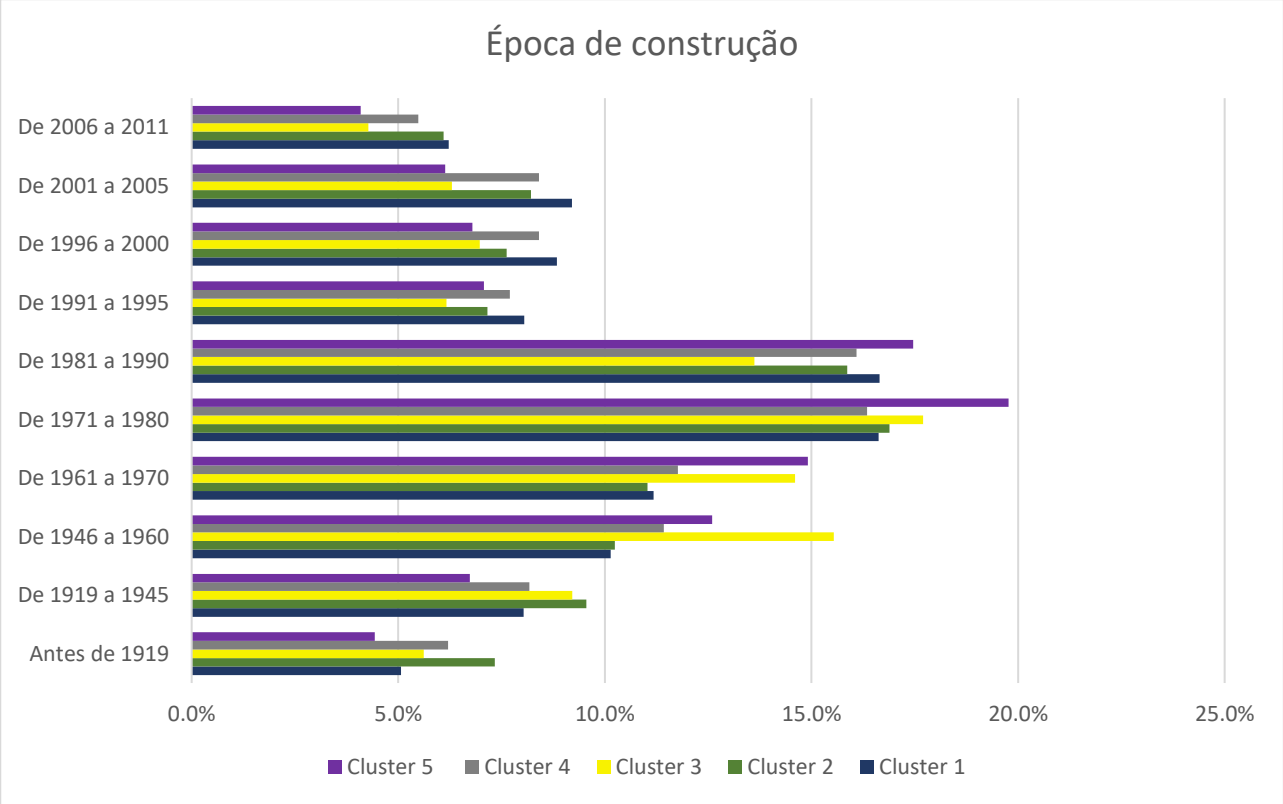

Figura 10 - Construção de edifícios por espaço temporal

Ao observar a época de construção de edifícios verifica-se uma predominância do *cluster* 2 de 1919 a 1945, com crescimento notável de 2006 a 2011. Posteriormente verifica-se um aumento notável do *cluster* 3 de 1946 a 1960 e do *cluster* 5 de 1961 a 1990. De 1991 a 2011, o *cluster* 1 domina a percentagem de edifícios construídos.

3.6.7 - Gênero por idade

Esta informação refere-se à distribuição da população em 5 faixas etária por gênero, durante uma data especifica.

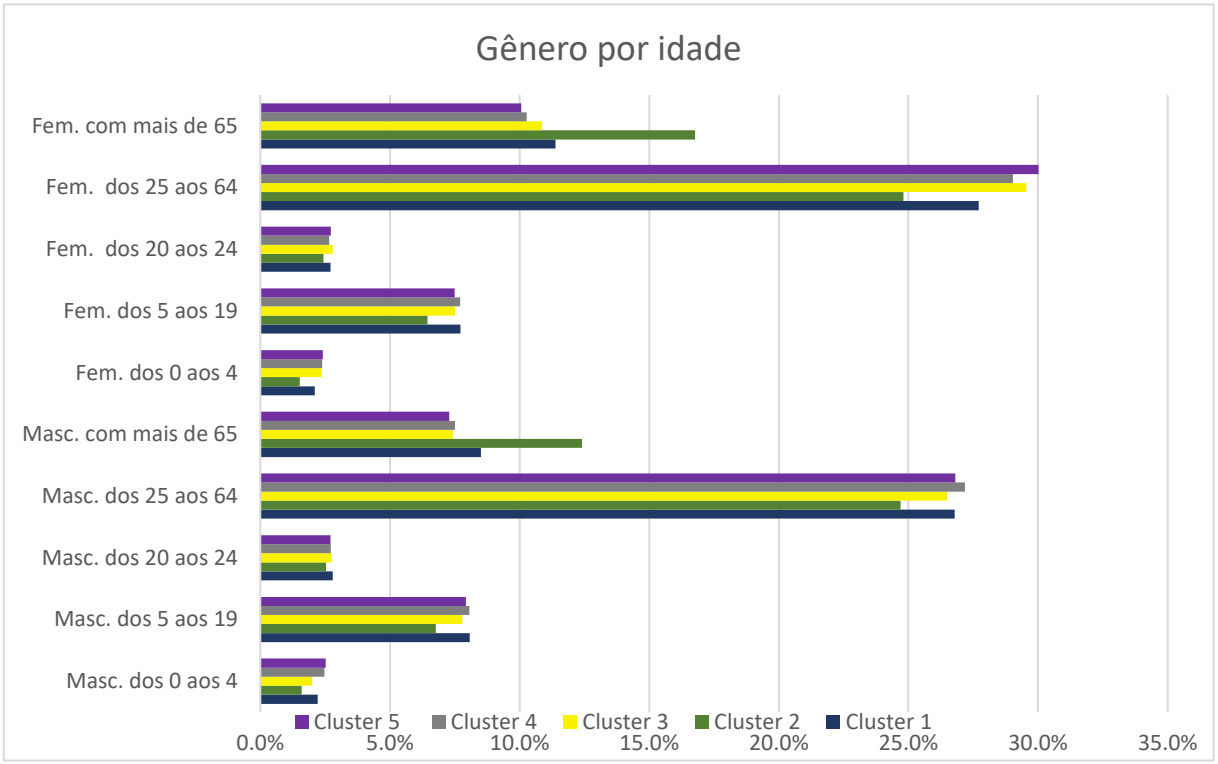

Figura 11 - Distribuição de idades por gênero

Analisando a distribuição de idades, podemos observar que o *cluster* 2 possui uma população muito mais envelhecida que os restantes, com um elevado número de indivíduos com mais de 65 anos e um número de indivíduos jovens abaixo da média. Os *clusters* 4 e 5 apresentam uma população com predominância em indivíduos de ambos os géneros 25 aos 64 anos e uma percentagem de crianças dos 0 aos 4 anos acima da média.

3.6.8 - Quintis de rendimento

Os quintis permitem ordenar, em partes iguais, 5 diferentes conjuntos de rendimentos. Para Portugal os quintis estão divididos da seguinte maneira: 1º quintil: menos de 13,296€ / ano; 2º quintil: entre 13,296€ e 19,250€ / ano; 3º quintil: entre 19,250€ e 26,418€ / ano; 4º quintil: entre 26,418€ e 38,608€ / ano; 5º quintil: 38,608€ / ano ou superior.

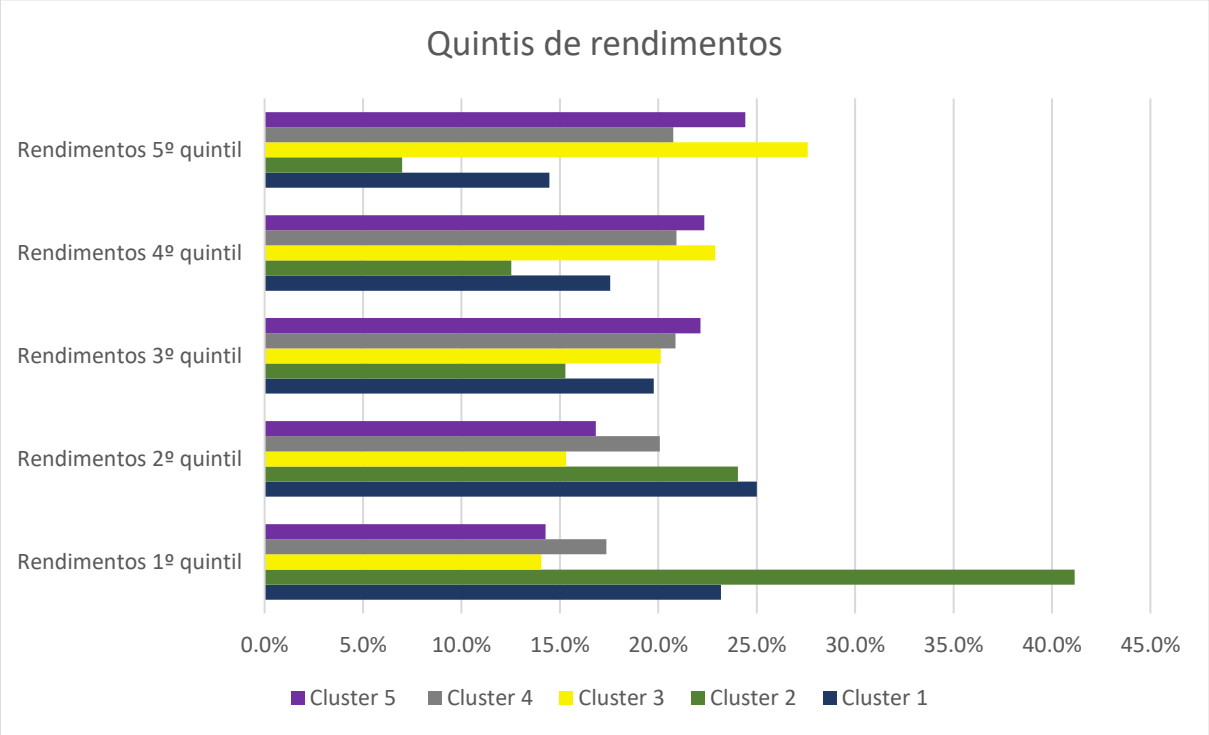

Figura 12 - Quintis de rendimentos da população

Analisando os quintis de rendimentos, é possível verificar que a maioria dos indivíduos do *cluster* 2 estão inseridos no primeiros quintis, com uma grande predominância do 1º quintil, sendo o agrupamento com menores rendimentos. O *cluster* 1 também apresenta baixos rendimentos para a maioria da população, com destaque para o 2º quintil. Contrariamente, os *clusters* 3 e 5 apresentam a maior percentagem de indivíduos inseridos nos últimos quintis, com uma predominância nos 4º e 5º quintis e no 3º e 4º quintis respetivamente. É ainda interessante verificar uma distribuição equilibrada entre os quintis do 4º *cluster*.

3.6.9 - Densidade populacional

A densidade populacional é um fator importante visto refletir a quantidade de indivíduos presentes por unidade de terra.

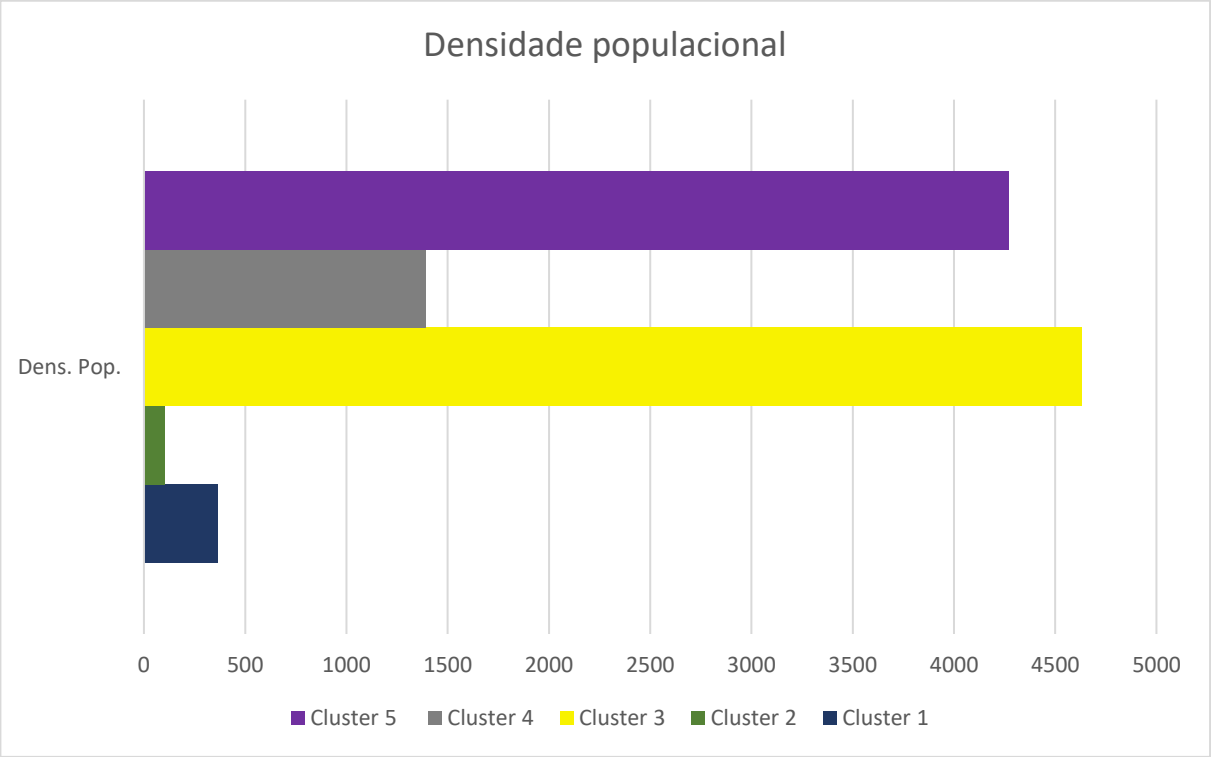

Figura 13 - Densidade populacional

Comparando os agrupamentos em termos da sua densidade populacional, observamos que os *clusters* 3 e 5 possuem um elevado número de indivíduos por unidade de área comparativamente aos restantes. Contrariamente, o *cluster* 2 apresenta uma densidade populacional muito reduzida.

3.6.10 - Descrição geral dos *clusters*

De acordo com a análise individual de cada variável por *cluster* pretende-se neste subcapítulo apresentar uma descrição geral das variáveis mais contributivas para cada *cluster* de forma a facilitar a caracterização dos diferentes *clusters*. São propostas 5 descrições gerais:

Cluster 1: descreve freguesias com uma forte presença industrial e baixa densidade populacional.

As famílias albergam maioritariamente habitações próprias, espaçosas, com bastantes divisões, muitas com 3 ou 4 indivíduos, apresentando um considerável número de indivíduos jovens e adultos.

Relativamente ao emprego, existe um grande foco no setor secundário, com alguma relevância do setor primário. Os rendimentos em geral são baixos, com os três primeiros quintis de rendimentos bastante representativos.

Salienta-se o crescimento substancial da construção de edificado de 1991 a 2011.

Exemplos: Monchique, Odemira, Aljustrel, Nisa, Mêda.

Cluster 2: cobre a grande maioria do interior de Portugal, com muito baixa densidade populacional, provavelmente devido ao êxodo rural.

Os agregados familiares habitam residências próprias, com muitas divisões e tamanho acima da média. Na sua maioria as famílias são envelhecidas, apresentando um elevado número de indivíduos com idades acima dos 65 anos e um baixo número de população jovem e casais com filhos.

Com um elevado número de reformados e trabalhadores domésticos, o setor primário e secundário emprega muita da população. A maioria dos trabalhadores encontra-se nos 1º e 2º quintis de rendimentos, com o 5º quintil muito pouco representado.

É possível ainda verificar um crescimento da construção de edifícios de 2005 a 2011 e a baixa escolaridade da população com um elevado número de analfabetos.

Exemplos: Ferragudo, Alijó, Murça.

Cluster 3: é composto por freguesias com uma densidade populacional, muito elevada e maior concentração de indivíduos com salários elevados.

As famílias são maioritariamente compostas por 1 ou 2 pessoas, vivem em habitações de tamanho abaixo da média, existindo algumas famílias a morar em habitações muito pequenas (inferior a 50 m^2). A população em geral é jovem adulta, com um baixo número de indivíduos com mais de 65 anos. De notar ainda um elevado número de residências arrendadas.

Com o setor terciário como grande empregador da população e uma baixa taxa de analfabetismo, mais de metade da população pertence aos 4º e 5º quintis de rendimentos, sendo assim as freguesias que apresentam maior poder monetário.

Com a construção de edifícios muito acentuada de 1946 a 1980, as freguesias deste agrupamento parecem representar os grandes centros urbanos, normalmente capitais de distrito, onde estão sediados a maioria dos grupos empresariais, centros de serviços públicos e grandes mercados.

Exemplos: Lisboa, Santarém, Lagos, Braga.

Cluster 4: os municípios pertencentes ao *cluster* 4 demonstram uma densidade populacional com valores próximos da média, sendo dissimilar dos outros clusters por apresentar percentagens medianas em praticamente todos os atributos.

Em grande parte, as famílias são compostas por 1 ou 2 indivíduos, apesar de se verificar um número notável de famílias com 3 ou 4 indivíduos, muitas delas com filhos, residindo em habitações de média dimensão. De notar um número considerável de habitações arrendadas.

Os indivíduos, em geral, trabalham no setor terciário com alguma representatividade do sector secundário, estando os 4 últimos quintis de rendimento distribuídos de igual forma.

De notar uma grande quantidade da população em idade ativa (dos 25 aos 64 anos) assim como bastantes indivíduos dos 0 aos 19 anos, um número baixo de reformados e um aumento da construção de edifícios de 1991 a 2005.

Exemplos: Sines, Beja, Lamego, Chaves.

Cluster 5: as freguesias do Cluster 5 albergam um elevado índice de população e encontram-se maioritariamente no litoral de Portugal.

Os agregados familiares são maioritariamente compostos por 1 ou 2 elementos, com grande representatividade da população em idade ativa (25 aos 64 anos). Estes vivem em habitações de dimensão média-baixa verificando-se um elevado número de habitações alugadas.

Nestas freguesias verifica-se uma percentagem significativa da população empregada no setor terciário, com baixo número de reformados, e uma percentagem considerável de desempregados. A população pertence na sua grande parte ao 3º, 4º e 5º escalões de rendimentos, com uma elevada percentagem no 5º quintil e uma percentagem acima da média no 3º quintil.

Estas freguesias estão em geral perto da costa, que, juntando aos baixos níveis de indústria primária e secundária poderá indicar que a maioria destas freguesias possui características propícias para o turismo.

Exemplos: Estoril, Portimão, Setúbal, Matosinhos.

4 – Validação dos resultados

Uma vez definida a segmentação, é necessário analisar os resultados de maneira a verificar a fiabilidade da proposta. Para este fim, utilizou-se um subconjunto de dados demonstrativos, com informação sobre o gasto total em compras e o gasto parcial em produtos frescos por freguesia, com o objetivo de compreender se a segmentação proposta se enquadra com o padrão espacial demonstrado. Para compreender o padrão espacial dos atributos relativos à informação usada foi realizada uma análise espacial de autocorrelação usando o *Moran* local. O *Moran* local é uma metodologia estatística proposta por *Anselin* (1995) para identificar padrões espaciais (por exemplo *clusters* e *outliers*) de uma dada variável. Para cada unidade espacial, este indicador avalia a sua significância estatística, e é obtido da seguinte maneira (equação 6) (*ESRI*, 2005):

$$I_i = \frac{x_{i-\bar{X}}}{S_i^2} \sum_{j=1, j \neq 1} w_{i,j} (x_j - \bar{X}) \qquad (6)$$

Onde:

x_i – observação *i* da variável em estudo;

$\bar{X}$ – média do atributo correspondente;

$w_{i,j}$ – Matriz de vizinhança espacial i e *j;*

Onde S representa o desvio padrão (equação 7):

$$S_i^2 = \frac{\sum_{j=1, j \neq 1}^{n} (x_j - \bar{X})^2}{n-1} \qquad (7)$$

Valores positivos para *I* indicam que a unidade de área sob analise se encontra localizada numa vizinhança com outras áreas de valores da variável similares, quer sejam altos ou baixos, indicando que essa *feature* (objeto do mundo real representado num mapa) pertence a um *cluster*. Um valor negativo indica que a *feature* tem outras *features* na sua vizinhança com valores dissimilares, sendo considerado um *outlier*. Em qualquer um dos casos, o seu nível de significância tem que ser pequeno o suficiente para que o *cluster* ou o *outlier* seja considerado estatisticamente significante.

Quando classificados como *cluster* ou *outlier*, as observações significativamente estatísticas são distinguidas como *clusters* de valores altos (HH), *clusters* de valores baixos (LL), *outliers* onde um valor alto está rodeado essencialmente por valores baixos (HL), e *outliers* onde um valor baixo está rodeado por um valor alto (LH).

Os dados utilizados para esta análise, o gasto total em compras e o gasto parcial em produtos frescos, contêm a sua localização e estão agrupados por freguesia, sendo que a média destes atributos foi calculada e utilizada como entrada na ferramenta *Cluster and outlier analysis (Anselin Local Moran's I*) (http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/cluster-and-outlier-analysis-anselin-local-moran-s.htm*)*, disponível no *ArcGIS Pro*.

Esta ferramenta necessita de três parâmetros obrigatórios de entrada: critério de vizinhança, medida de distância e normalização. O critério de vizinhança irá indicar como será feita a identificação de vizinhos de cada *feature*, definindo a relação espacial entre eles. A medida de distância especifica como são calculadas as distâncias de cada *feature* a *features* consideradas vizinhas. Para o desenvolvimento deste projeto foi utilizada a distância inversa como critério de vizinhança (*features* vizinhas têm maior influência no cálculo do índice, que *features* distantes), a distância euclidiana como parâmetro de distância e normalização por linha.

Algumas freguesias continham poucas observações disponíveis, tendo sido mantidas apesar disso, devido à importância e expressividade que os dados introduziam no projeto. Estes dados pertencem a freguesias do distrito de Lisboa e Setúbal.

4.1. Freguesias e *clusters* correspondentes

Das 2882 freguesias existentes em Portugal Continental, 135 continham informação sobre consumos. De maneira a melhorar a legibilidade da cartografia, as freguesias foram numeradas de 1 a 135, estando a listagem correspondente e uma aproximação das freguesias centrais disponível no anexo 2 e 4 respetivamente. A Figura 14 representa as freguesias que continham informação sobre o consumo total em compras e consumo parcial de produtos frescos.

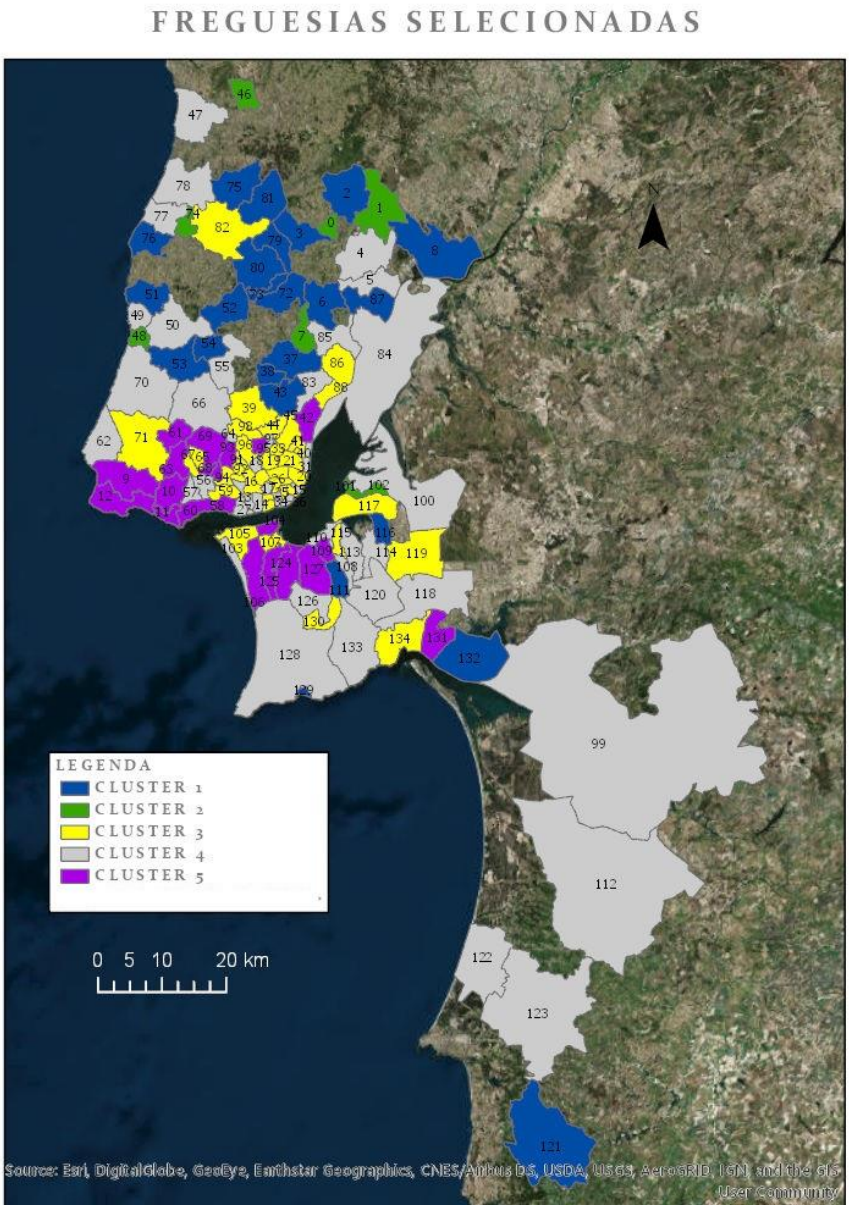

Figura 14 - Freguesias com informação relativa a gastos

A distribuição freguesias selecionadas por *clusters* é apresentada na tabela 6:

Tabela 6 - Distribuição de freguesias selecionadas por cluster

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Nº de freguesias	24	8	38	45	20
Percentagem (%)	17.8	5.9	28.2	33.3	14.8

4.2 *Moran* local – Gastos total em compras

Visto que os dados em analise são referentes a consumos podemos assumir que, quanto maior o gasto total, maior será o poder de compra de uma determinada freguesia. Desta forma podemos comparar os resultados obtidos com as descrições anteriormente feitas, que consideram diferentes níveis económicos para cada *cluster*. Recordando que, em termos de poder monetário, os *clusters* podem ser ordenados de maneira crescente da seguinte maneira: 2, 1, 4, 5, 3.

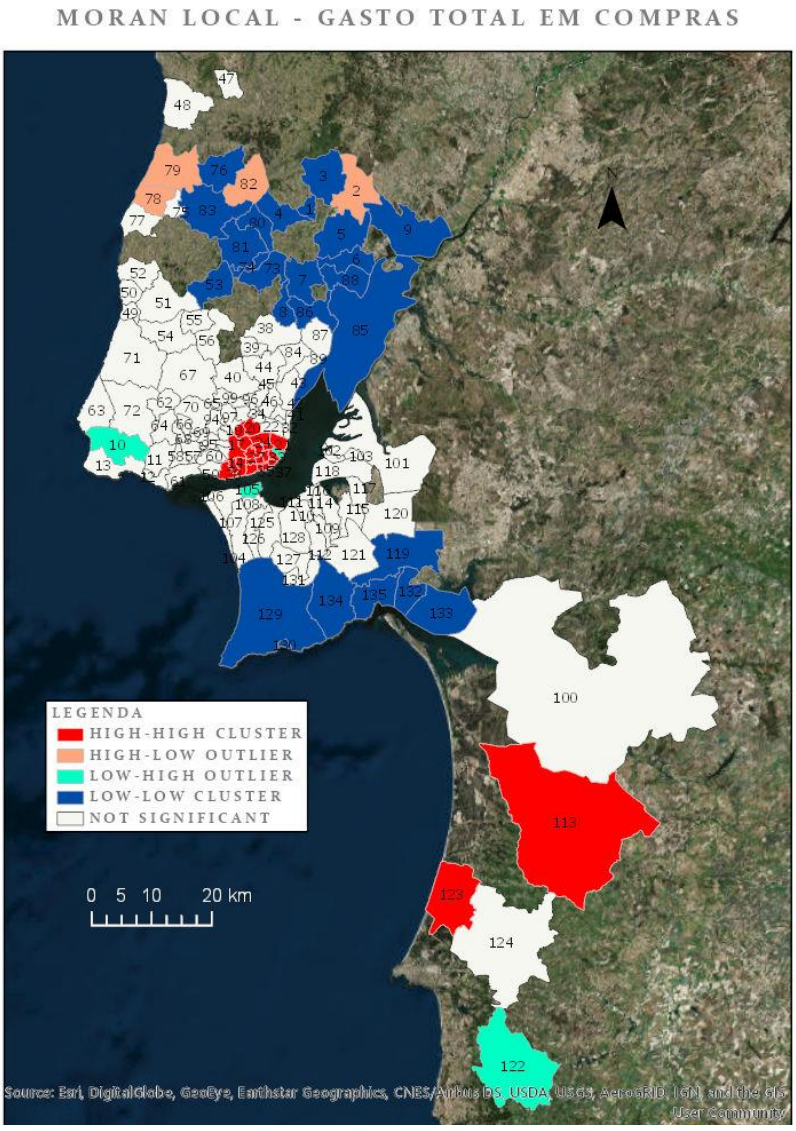

Figura 15 - Moran local do consumo de produtos diversos

Calculado o índice local de *Moran*, obtemos um mapa indicador de autocorrelação espacial. Como mencionado anteriormente, as observações significantemente estatísticas podem ser distinguidas como: observação e vizinhança com valores altos (HH); observação e vizinhança

com valores baixos (LL); observação com valor alto, mas vizinhança com valores baixos (HL); observação com valor baixo, mas vizinhança com valores altos (LH).

Podemos observar que apenas as freguesias de Lisboa apresentam valores altos rodeados de valores altos (HH) para gastos totais em produtos, e a existência de dois agrupamentos com valores baixos e rodeados de valores baixos (LL). Verifica-se ainda algumas freguesias isoladas (LH e HL) que não apresentam um padrão claro de valores de gastos na sua vizinhança. Este mapa encontra-se sumarizado na tabela 7.

Tabela 7 - Classificação do Moran local para produtos diversos por cluster

Moran's I	Cluster	Freguesias	Total
HH	3	{17,20,21,23,24,25,26,27,29,30,33}	$HH_3 = 11$
	4	{14,15,18, 19,28,31,35,36,37,113,123}	$HH_4 = 11$
HL	1	{82}	$HL_1 = 1$
	2	{2}	$HL_2 = 1$
	4	{78,79}	$HL_4 = 2$
LH	1	{122}	$LH_1 = 1$
	4	{16}	$LH_4 = 1$
	5	{10,105}	$LH_5 = 2$
LL	1	{3,4,7,9,53,73,74,76,80,81,88,130,133}	$LL_1 = 13$
	2	{1,8}	$LL_2 = 2$
	3	{83,135}	$LL_3 = 2$
	4	{5,6,85,86,119,129,134}	$LL_4 = 7$
	5	{132}	$LL_5 = 1$

Das 135 freguesias utilizadas, 55 demonstram significância estatística, tendo sido classificados como *clusters* ou *outliers*. Comparando os resultados da segmentação com os resultados do *Moran* local observamos que:

- O HH (valor e vizinhança com valores altos) é composto por 22 freguesias repartidas igualmente pelos *clusters* 3 e 4. Segundo a descrição de ambos, o 3 *cluster* enquadra-se nos resultados visto ser o *cluster* com maior poder de compra. O *cluster* 4 poderá enquadrar-se, visto uma percentagem considerável dos indivíduos deste agrupamento pertencer aos quintis de rendimentos mais altos, aliado ao fato de ser espectável que exista maior poder monetário na área da Grande Lisboa quando comparado ao restante Portugal;
- Os HL (valor alto, mas vizinhança com valores baixos) contêm $HL_1 = 1$, $HL_2 = 1$ e $HL_4 = 2$. De acordo com a segmentação proposta, os *clusters* 1 e 2 apresentam os menores valores de rendimentos como indicado pelo resultado. Relativamente ao *cluster* 4, podemos observar que as freguesias consideradas *outliers* encontram-se em zona de fronteira, o que piora a fiabilidade dos resultados. Em todo o caso, são esperados valores discrepantes entre observações do *cluster* 4 a nível de rendimentos visto existir uma distribuição equilibrada nos rendimentos da população, como demonstrado anteriormente;
- Os LH (valor baixo, mas vizinhança com valores altos) identificados são $LH_1 = 1$, $LH_4 = 1$, $LH_5 = 2$. Verifica-se esta situação, segundo a segmentação efetuada, nos *clusters* 2 e 4, principalmente tendo em conta que as freguesias consideradas pertencem a zona central de Lisboa. A freguesia pertencente ao *cluster* 1 por sua vez apresenta apenas um vizinho, numa zona de fronteira, o que reduz a fiabilidade do resultado.

- O LL (valor baixo e vizinhança com valores baixos) contem freguesias de todos os *clusters*: $LL_1 = 13$; $LL_2 = 2$; $LL_3 = 2$; $LL_4 = 7$; $LL_5 = 1$. De acordo com a segmentação elaborada, as freguesias do *cluster* 1 e 2 enquadram-se nestes resultados. Contrariamente, 3 das observações consideradas como baixas em termos de consumos foram identificadas como freguesias com elevado poder monetário (*clusters* 3 e 5).

4.3. *Moran* local – Gastos parciais em produtos frescos

Quando calculado o índice de *Moran* local da informação sobre gastos parciais em produtos frescos obtemos resultados semelhantes aos anteriores (Figura 16):

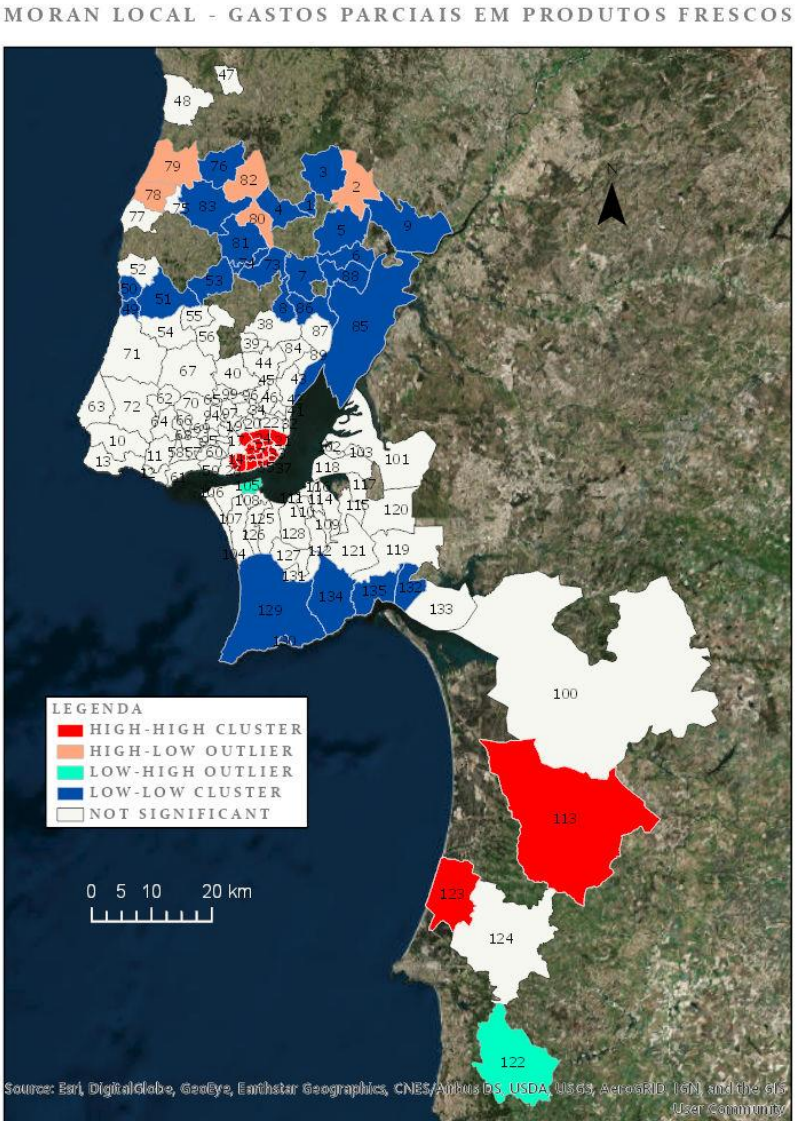

Figura 16 - Moran local de gastos parciais em produtos frescos

Identificadas as 50 freguesias com significância estatística, é possível cruzar estes resultados com a segmentação proposta (Tabela 8):

Tabela 8 - Classificação do Moran local para produtos frescos por cluster

Moran's I	Cluster	Freguesias	Total
HH	3	{21,23,24,25,26,27,29,30,31,33}	$HH_3 = 9$
	4	{14,15,18,31,35,36,37,113,123}	$HH_4 = 9$
HL	1	{80,82}	$HL_1 = 2$
	2	{2}	$HL_2 = 1$
	4	{78,79}	$HL_4 = 2$
LH	1	{122}	$LH_1 = 1$
	5	{105}	$LH_5 = 1$
LL	1	{3,4,7,9,53,73,74,76,81,88,130}	$LL_1 = 11$
	2	{1,8,49}	$LL_2 = 3$
	3	{83,135}	$LL_3 = 2$
	4	{5,6,50,51,85,86,129,134}	$LL_4 = 8$
	5	{132}	$LL_5 = 1$

Analisando os resultados é possível verificar que:

- Similarmente à primeira análise, as observações e vizinhança com valores altos (HH) é composto pelos *clusters 3* e 4. De notar uma redução das freguesias significantemente estatísticas;
- No caso dos HL, verifica-se o mesmo resultado obtido na análise anterior referente ao gasto total de compras;
- Os LH são compostos por freguesias do *cluster 1* e *5*. Quando comparado com os resultados do *Moran* local do gasto total de compras, verificamos que não foi considerada uma freguesia do *cluster 4* e uma do *cluster* 5.
- O LL é composto por freguesias de todos os agrupamentos, como verificado no LL da análise anterior. É possível verificar que as freguesias 80 e 133 foram removidas deste *cluster*, foi adicionada a freguesia 49 pertencente ao *cluster* 2 e foram removidas as freguesias 50 e 51 e adicionada a 119, pertencentes ao *cluster* 4.

4.4. Discussão dos resultados

Das 55 freguesias selecionadas como estatisticamente significantes, para o gasto total em produtos, 30 freguesias encontram-se em concordância e 4 diferem da segmentação proposta, as restantes 21 freguesias pertencem ao *cluster* 4, que como mencionado anteriormente, é esperado que contenha valores nos dois extremos. O Moran local para gastos parciais em produtos frescos apresenta 27 freguesias em conformidade com a designação proposta e 4 em discordância, com 20 freguesias pertencentes ao *cluster* 4.

Em suma, dos atributos avaliados em termos de autocorrelação espacial, verifica-se que a maioria das zonas os gastos são mais elevados (HH) correlacionam-se espacialmente com as freguesias pertencentes ao *cluster* 3, sendo este agrupamento o que apresenta maior poder de compra comparativamente aos restantes clusters. As zonas com gastos baixos (LL) apresentam uma correlação espacial com as freguesias do *cluster* 1 e 2, os quais apresentam os menores rendimentos, quando comparados com os restantes *clusters*. O *cluster* 4 encontra-se presente em todas as classes produzidas pelo *Moran* local, derivado do equilíbrio entre atributos verificados na descrição do mesmo. Desta forma, é possivel aceitar as variáveis utilizadas e a segmentação proposta, face à análise produzida de forma independente usando valores de gastos em produtos.

5 – Conclusões e trabalhos futuros

5.1. Conclusões

Neste projeto desenvolveu-se uma segmentação para as freguesias de Portugal Continental, baseada em características da população como o agregado familiar, condição face ao emprego, dimensões de alojamentos, nível de educação adquirido, idade por gênero, densidade populacional, quintis de rendimentos e época de construção de edifícios. As variáveis escolhidas como base para a segmentação foram selecionadas de acordo com um índice de validade interna, denominado *silhouette score*.

As freguesias foram segmentadas em 5 diferentes agrupamentos, denominados de *clusters*, onde cada *cluster* possuí diferentes características, refletindo os atributos das freguesias que os constituem. Para tal foi utilizado o método *k-means*, que particiona a informação de forma a minimizar a soma dos quadrados dos erros produzidos.

Dos agrupamentos resultantes, verificou-se que o *cluster* 2 continha 72% das freguesias selecionadas, estando estas maioritariamente localizadas no interior de Portugal. Estas são descritas como freguesias pobres e com pouca população residente, derivado da centralização da indústria e serviços em capitais de distrito. Os *clusters* 3 e 5 parecem agrupar essas capitais de distrito e cidades com elevada densidade populacional e capacidade monetária, normalmente próximas entre si e localizadas no litoral de Portugal. Estes *clusters* contêm, respetivamente, 2% e 1% das observações consideradas.

De maneira a validar a segmentação proposta foi utilizado um subconjunto de dados demonstrativos de consumo que continham informação sobre o gasto total de compras e o gasto parcial em produtos frescos. Para tal, foi calculado um índice local de autocorreção espacial para os dois conjuntos de dados independentes. Quando comparados com a segmentação, verificou-se que os resultados satisfaziam o nível de confiança necessário para a aceitação da cartografia.

A segmentação realizada demonstra potencial no sentido em que fornece uma caracterização da população permitindo abordar novas oportunidades, adaptar e melhorar a precisão e eficiência de campanhas de *marketing* adequando o produto ou serviço às necessidades do mercado-alvo, ser uma ferramenta base para estudos económicos e/ou sociais, ajustar produtos e canais de distribuição em função das necessidades dos segmentos entendidos como favoráveis a serem explorados comercialmente, adaptar o meio de comunicação e a publicidade produzida de maneira a reduzir custos, otimizar a localização de pontos de venda de acordo com os segmentos considerados de interesse, auxiliar na identificação de concorrência e suas vulnerabilidades, entre outros.

Por fim, considera-se que o maior problema deste projeto se prendeu com a falta de fontes de dados diversas. Quando se desenvolve uma segmentação é importante que a informação seja variada, de forma a atribuir uma maior expressividade aos *clusters* criados. Apesar da grande quantidade de informação disponibilizada nos censos de Portugal Continental, seria importante a utilização de dados que melhor refletissem a afluência e estilo de vida da população, de maneira a melhorar o particionamento e a diferenciação dos *clusters* propostos.

5.2. Trabalhos futuros

De maneira a melhorar estas análises, seria importante o desenvolvimento de bases de dados georreferenciadas, compostas por diferentes fontes de dados de maneira a melhorar estudos atuais e proporcionar o desenvolvimento de novos estudos. Seria também importante adaptar estas bases de dados a uma menor escala geográfica, visto muita da informação disponível para Portugal Continental se encontrar indisponível para a subsecção estatística.

Relativamente à metodologia aplicada, seria interessante a exploração de diferentes metodologias de *clustering*, como por exemplo mapas de *Kohonen*, e explorar mais

aprofundadamente cada *cluster*, no sentido de compreender se seria possível subdividir os *clusters* originais em novos segmentos, mais explicativos que os propostos.

6 – Referências bibliografias

- Armstrong, G. e P. Kotler (2005), Marketing: An Introduction. 7 rd ed. Upper Saddle River, N.J: Prentice Hall;
- Bailey, S., Charlton, J., Dollamore, G., Fitzpatrick, J. (2000) Families, groups and clusters of local and health authorities of Great Britain: revised for authorities in 1999. Popln Trends, 99, 37–52;
- Brown, PJB. (1991) Exploring geodemographics, in Masser I and Blakemore M (eds) Handling geographic information, Longman, London, 221-258;
- C. Liu, T. Hu, Y. Ge, H. Xiong (2012), Which Distance Metric is Right: An Evolutionary K-Means View, Sdm'12, pp. 907–918;
- CACI (2018) Acorn Technical Guide. URL: https://www.esri.com/library/whitepapers/pdfs/esri-data-tapestry-segmentation.pdf Acedido em (16 de Agosto de 2018)https://www.caci.co.uk/sites/default/files/resources/Acorn_technical_guide.pdf
- Charlton, M., Openshaw, S., and Wymer, C. (1985) Some new classifications of census Enumeration Districts in Britain: a poor man's ACORN, Journal of Economic and Social Measurement, 13, 69-96.
- Cheng, E.W.L., Li, H., Yu. L. (2007) A GIS Approach to Shopping Mall Location Selection. Building and Environment, Vol. 42, No. 2, pp 884-892;
- D. Vickers and P. Rees (2006), Creating the UK National Statistics 2001 output area classification, J. R. Stat. Soc. Ser. A Stat. Soc., vol. 170, no. 2, pp. 379–403;
- Debenham, J. (2002) Understanding Geodemographic Classification: Creating the Building Blocks For An Extension. Working Paper. School of Geography, University of Leeds;
- Dillion, J., Rickinson, M., Sanders, D.L., (2006) The value of outdoor learning: evidence from research in the UK and elsewhere,Sch. Sci. Rev., vol. 7, no. 320, pp. 107–112.
- Dudoit, S., Yang, Y. H., Callow, M. J., Speed, T. P., (2002) Statistical Methods For Identifying Differentially Expressed Genes In Replicated Cdna Microarray Experiments, vol. 12, pp. 111–139;
- ESRI (2014) Tapestry Segmentation: Methodology. URL: http://downloads.esri.com/esri_content_doc/dbl/us/J9941_Tapestry_Segmentation_Methodology_2018.pdf Acedido em (08 de Agosto de 2018)
- Experian (2014) Under the bonnet: Mosaic data, methodology and build. URL: https://docslide.us/documents/under-the-bonnet-mosaic-data-methodology-and-in-methodology-optimising-mosaic.html Acedido em (16 de Maio de 2018)
- Freire S. (2010) Geographic Information and Cartography for Risk and Crisis Management, no. January 2010;
- Gunter, Barrie, Furnham A. (1992) Consumer profiles: An introduction to psychographics Routledge, London;
- Han, Jiawei, Kamber, Micheline (2000). Data Mining: Concepts and Techniques. Morgan Kaufmann;
- Harris, R., Sleight, P., Webber, R. (2005) Geodemographics, GIS and Neighbourhood Targeting. Chichester:Wiley;
- Johnson, M.D., Gustafsson A. (2000), Improving Customer Satisfaction, Loyalty e Profit: An Integrated Measurement e Management System (J-B-UMBS Series), CA Journal of Marketing 28 (10), 49-66;
- K. Jain, (2010) Data clustering: 50 years beyond K-means, Pattern Recognit. Lett., vol. 31, no. 8, pp. 651–666;
- Karthikeyani, V.N., Thangavel, K., (2009) Impact of normalization in distributed K-means clustering. Int. J. Soft Compute., 4(4): 168-172;
- Kaufman, L., Rousseauw, P.J. (1990), Finding groups in data. An introduction to cluster analysis, Wiley, New York;

- Kotler, P. (1984) Marketing Management. Administração de Marketing – Análise, Planejamento, Implementação e Controle. Editora Atlas, 5a edição, São Paulo;
- Kotler, P. & Keller, K. L. (2009) Marketing Management. Pearson Education International, 13. Edition;
- Lletí, R., Ortiz, M. C., Sarabia, L. A., Sánchez, M. S., (2004) Selecting variables for k-means cluster analysis by using a genetic algorithm that optimizes the silhouettes, Anal. Chim. Acta, vol. 515, no. 1, pp. 87–100;
- Mishra, S. (2009) GIS in Indian retail industry – a strategic tool, International Journal of Marketing Studies, Vol. 1, No.1, pp 50-57;
- Mohamad, B., Usman, D., (2013) Standardization and its effects on K-means clustering algorithm, Res. J. Appl. Sci. Eng. Technol., vol. 6, no. 17, pp. 3299–3303.
- Pickton, D. e Broderick, A. (2005), Integrated marketing communications. 2.edition: Financial Times/ Prentice Hall;
- Salvador S., Chan, P. (2003) Determining the Number of Clusters/Segments in Hierarchical Clustering/Segmentation Algorithms, Dept. of Computer Sciences Technical Report CS-2003-18;
- Sleight, P. (1997) Targeting Customers: How to Use Geodemographic and Lifestyle Data in Your Business, NTC Publications, Henley-on-Thames;
- Sun, S. (2009), An Analysis on the Conditions and Methods of Market Segmentation, International Journal of Business e Management 4 (2), 63-69;
- Theodoridis, S., Koutroumbas, K. (2008) Pattern Recognition, 4th Edition
- Tou, J. T., & Gonzalez, R. C. (1974). Pattern recognition principles. London: AddisonWesley;
- Tseng, Y. Y., Ubbels, B., Verhoef, E., (2005) Value of time, schedule delay, and reliability – Estimation results of a stated choice experiment among Dutch commuters facing congestion. Paper presented at the 45th Congress of European Regional Science Association;
- Vaishali, R.P., Rupa, G.M., (2011). Impact of outlier removal and normalization approach in modified k-means clustering algorithm. Int. J. Comput. Sci., 8(5): 331-336;
- Webber, R. (2004). Designing Geodemographics systems to meet contemporary business needs, Interactive Marketing;
- Weinstein, A. (2004) Handbook of market segmentation: strategic targeting for business and technology firms. New York: Routledge.
- Wendel, M, W.A. Kamakura (1998), Market Segmentation: Conceptual e Methodological Foundations. Boston MA. Kluwer Academic Publishers
- Wendel, M. e W.A. Kamakura, (2000), *Market Segmentation: Conceptual e Methodological Foundations.* Boston MA. Kluwer Academic Publishers.
- Wendell, R. S. (1956). Product differentiation and market segmentation as alternative marketing strategies. Journal of Marketing, 21, 3–8;

Anexos

ANEXO 1 – Descrição das variáveis

i. Agregado familiar – conjuga informação sobre a dimensão de uma família e a sua condição perante a habitação:
    a. 1 ou 2 indivíduos – Famílias com 1 ou 2 indivíduos;
    b. 3 ou 4 indivíduos – Famílias com 3 ou 4 indivíduos;
    c. 5 ou mais indivíduos – Famílias com mais de 5 indivíduos;
    d. Núcleos familiares com filhos – Núcleos familiares com filhos;
    e. Núcleos familiares sem filhos – Núcleos familiares sem filhos;
    f. Residência com proprietário ocupante – Ocupante da habitação é o proprietário;
    g. Residência alugada – Habitação alugada por outrem.

ii. Emprego – refere-se aos diferentes sectores que empregam a população e ocupação de não trabalhadores:
    a. Sector primário – Empregados do sector primário;
    b. Sector secundário – Empregados do sector secundário;
    c. Sector terciário – Empregados do sector terciário;
    d. Reformados – Individuo reformado;
    e. Trabalhadores domésticos – Individuo não está empregado, mas também não se encontra à procura emprego;
    f. Desempregados – Individuo desempregado, mas à procura.

iii. Divisões dos alojamentos – Número de divisões por alojamento:
    a. 1 ou 2 divisões – Habitação com 1 ou 2 divisões;
    b. 3 ou 4 divisões – Habitação com 3 ou 4 divisões;
    c. Mais de 4 divisões – Habitação com mais de 4 divisões;

iv. Dimensões dos alojamentos – Dimensão dos alojamentos:
    a. Menos de 50 m^2 – Alojamentos com menos de 50 m^2;
    b. Entre 50 e 100 m^2 – Alojamentos entre 50 e 100 m^2;
    c. Entre 100 e 200 m^2 – Alojamentos entre 100 e 200 m^2;
    d. Mais de 200 m^2 – Alojamentos com mais de 200 m^2.

v. Educação – Nível mais alto de educação completado por um individuo:
    a. Básico – Ensino básico;
    b. Secundário – Ensino secundário;
    c. Superior – Ensino pós-básico ou superior;
    d. Analfabetos – Indivíduos que não sabem ler nem escrever.

vi. Época de construção – indica o ano no qual a habitação foi construída, ou a idade da habitação:
    a. Antes de 1919 – Habitações de 1919 e antes;
    b. De 1919 a 1945 – Habitações construídas entre 1919 e 1945;
    c. De 1946 a 1960 – Habitações construídas entre 1946 e 1960;
    d. De 1961 a 1970 – Habitações construídas entre 1961 e 1970;
    e. De 1971 a 1980 – Habitações construídas entre 1971 e 1980;
    f. De 1981 a 1990 – Habitações construídas entre 1981 e 1990;
    g. De 1991 a 1995 – Habitações construídas entre 1991 e 1995;
    h. De 1996 a 2000 – Habitações construídas entre 1996 e 2000;
    i. De 2001 a 2005 – Habitações construídas entre 2000 e 2005;
    j. De 2006 a 2011 – Habitações construídas entre 2006 e 2011.

vii. Sexo por idade: - Idade de um indivíduo de um sexo específico, durante uma data especifica:
    a. Masculino dos 0 aos 4 – Homem com idade entre os 0 e os 4 anos;
    b. Masculino dos 5 aos 19 – Homem com idade entre os 5 e os 19 anos;
    c. Masculino dos 20 aos 24 – Homem com idade entre os 20 e os 24 anos;
    d. Masculino dos 25 aos 64 – Homem com idade entre os 25 e os 64 anos;
    e. Masculino com mais de 65 – Homem com idade superior aos 65 anos.
    f. Feminino dos 0 aos 4 – Mulher com idade entre os 0 e os 4 anos;
    g. Feminino dos 5 aos 19 – Mulher com idade entre os 5 e os 19 anos;
    h. Feminino dos 20 aos 24 – Mulher com idade entre os 20 e os 24 anos;
    i. Feminino dos 25 aos 64 – Mulher com idade entre os 25 e os 64 anos;
    j. Feminino com mais de 65 – Mulher com idade superior aos 65 anos.

viii. Quintis de rendimentos – Conjunto de rendimentos ordenados em cinco partes iguais:
    a. 1º quintil – Indivíduos com rendimentos anuais a baixo dos €13,296;
    b. 2º quintil – Indivíduos com rendimentos anuais entre os €13,296 e €19,250;
    c. 3º quintil – Indivíduos com rendimentos anuais entre os €19,250 e €26,418;
    d. 4º quintil – Indivíduos com rendimentos anuais entre os €26,418 e €38,608;
    e. 5º quintil – Indivíduos com rendimentos anuais acima dos €38,608.

ix. Densidade populacional – Número de indivíduos por unidade de área.

ANEXO 2 – Numeração das freguesias utilizadas na validação

ID	Freguesia
1	Olhalvo
2	Ota
3	União das freguesias de Abrigada e Cabanas de Torres
4	União das freguesias de Aldeia Galega da Merceana e Aldeia Gavinha
5	União das freguesias de Alenquer (Santo Estêvão e Triana)
6	União das freguesias de Carregado e Cadafais
7	Arruda dos Vinhos
8	S. Tiago dos Velhos
9	Azambuja
10	Alcabideche
11	São Domingos de Rana
12	União das freguesias de Carcavelos e Parede
13	União das freguesias de Cascais e Estoril
14	Ajuda
15	Alcântara
16	Beato
17	Benfica
18	Campolide
19	Carnide
20	Lumiar
21	Marvila
22	Olivais
23	São Domingos de Benfica
24	Alvalade
25	Areeiro
26	Arroios
27	Avenidas Novas
28	Belém
29	Campo de Ourique
30	Estrela
31	Misericórdia
32	Parque das Nações
33	Penha de França
34	Santa Clara
35	Santa Maria Maior
36	Santo António
37	São Vicente
38	Bucelas
39	Fanhões
40	Loures
41	União das freguesias de Moscavide e Portela
42	União das freguesias de Sacavém e Prior Velho
43	União das freguesias de Santa Iria de Azoia, São João da Talha e Bobadela
44	União das freguesias de Santo Antão e São Julião do Tojal
45	União das freguesias de Santo António dos Cavaleiros e Frielas
46	União das freguesias de Camarate, Unhos e Apelação
47	Reguengo Grande
48	União das freguesias de Lourinhã e Atalaia

49	Carvoeira
50	Ericeira
51	Mafra
52	Santo Isidoro
53	União das freguesias de Enxara do Bispo, Gradil e Vila Franca do Rosário
54	União das freguesias de Igreja Nova e Cheleiros
55	União das freguesias de Malveira e São Miguel de Alcainça
56	União das freguesias de Venda do Pinheiro e Santo Estêvão das Galés
57	Barcarena
58	Porto Salvo
59	União das freguesias de Algés, Linda-a-Velha e Cruz Quebrada-Dafundo
60	União das freguesias de Carnaxide e Queijas
61	União das freguesias de Oeiras e São Julião da Barra, Paço de Arcos e Caxias
62	Algueirão-Mem Martins
63	Colares
64	Rio de Mouro
65	Casal de Cambra
66	União das freguesias de Agualva e Mira-Sintra
67	União das freguesias de Almargem do Bispo, Pêro Pinheiro e Montelavar
68	União das freguesias do Cacém e São Marcos
69	União das freguesias de Massamá e Monte Abraão
70	União das freguesias de Queluz e Belas
71	União das freguesias de São João das Lampas e Terrugem
72	União das freguesias de Sintra (Santa Maria e São Miguel, São Martinho e São Pedro de Penaferrim)
73	Santo Quintino
74	Sobral de Monte Agraço
75	Ponte do Rol
76	Ramalhal
77	São Pedro da Cadeira
78	Silveira
79	União das freguesias de A dos Cunhados e Maceira
80	União das freguesias de Carvoeira e Carmões
81	União das freguesias de Dois Portos e Runa
82	União das freguesias de Maxial e Monte Redondo
83	Santa Maria, São Pedro e Matacães
84	Vialonga
85	Vila Franca de Xira
86	União das freguesias de Alhandra, São João dos Montes e Calhandriz
87	União das freguesias de Alverca do Ribatejo e Sobralinho
88	União das freguesias de Castanheira do Ribatejo e Cachoeiras
89	União das freguesias de Póvoa de Santa Iria e Forte da Casa
90	Alfragide
91	Águas Livres
92	Encosta do Sol
93	Falagueira-Venda Nova
94	Mina de Água
95	Venteira
96	Odivelas
97	União das freguesias de Pontinha e Famões
98	União das freguesias de Póvoa de Santo Adrião e Olival de Basto

99	União das freguesias de Ramada e Caneças
100	União das freguesias de Alcácer do Sal (Santa Maria do Castelo e Santiago) e Santa Susana
101	Alcochete
102	Samouco
103	São Francisco
104	Costa da Caparica
105	União das freguesias de Almada, Cova da Piedade, Pragal e Cacilhas
106	União das freguesias de Caparica e Trafaria
107	União das freguesias de Charneca de Caparica e Sobreda
108	União das freguesias de Laranjeiro e Feijó
109	Santo António da Charneca
110	União das freguesias de Alto do Seixalinho, Santo André e Verderena
111	União das freguesias de Barreiro e Lavradio
112	União das freguesias de Palhais e Coina
113	União das freguesias de Grândola e Santa Margarida da Serra
114	Alhos Vedros
115	Moita
116	União das freguesias de Baixa da Banheira e Vale da Amoreira
117	Sarilhos Grandes
118	União das freguesias de Montijo e Afonsoeiro
119	Palmela
120	Pinhal Novo
121	Quinta do Anjo
122	Cercal
123	Santo André
124	União das freguesias de Santiago do Cacém, Santa Cruz e São Bartolomeu da Serra
125	Amora
126	Corroios
127	Fernão Ferro
128	União das freguesias do Seixal, Arrentela e Aldeia de Paio Pires
129	Sesimbra (Castelo)
130	Sesimbra (Santiago)
131	Quinta do Conde
132	Setúbal (São Sebastião)
133	Sado
134	União das freguesias de Azeitão (São Lourenço e São Simão)
135	União das freguesias de Setúbal (São Julião, Nossa Senhora da Anunciada e Santa Maria da Graça)

Anexo 3 – Variáveis recolhidas

Variáveis	Descrição
n_edificios_classicos	Edifícios classicos
n_edificios_classicos_1ou2	Edificios clássicos construídos estruturalmente p/ possuir 1 ou 2 alojamentos
n_edificios_classicos_isolados	Edificios clássicos isolados
n_edificios_classicos_gemin	Edificios clássicos geminados
n_edificios_classicos_embanda	Edificios clássicos em banda
n_edificios_classicos_3oumais	Edificios clássicos construídos estruturalmente p/ possuir 3 ou mais alojamentos
n_edificios_classicos_outros	Outro tipo de edificio clássico
n_edificios_exclusiv_resid	Edifícios exclusivamente residenciais
n_edificios_principal_resid	Edifícios principalmente residenciais
n_edificios_princip_nao_resid	Edifícios principalmente não residenciais
n_edificios_1ou2_pisos	Edifícios com 1 ou 2 pisos
n_edificios_3ou4_pisos	Edifícios com 3 ou 4 pisos
n_edificios_5ou_mais_pisos	Edifícios com 5 ou mais pisos
n_edificios_constr_antes_1919	Edifícios construídos antes de 1919
n_edificios_constr_1919a1945	Edifícios construídos entre 1919 e 1945
n_edificios_constr_1946a1960	Edifícios construídos entre 1946 e 1960
n_edificios_constr_1961a1970	Edifícios construídos entre 1961 e 1970
n_edificios_constr_1971a1980	Edifícios construídos entre 1971 e 1980
n_edificios_constr_1981a1990	Edifícios construídos entre 1981 e 1990
n_edificios_constr_1991a1995	Edifícios construídos entre 1991 e 1995
n_edificios_constr_1996a2000	Edifícios construídos entre 1996 e 2000
n_edificios_constr_2001a2005	Edifícios construídos entre 2001 e 2005
n_edificios_constr_2006a2011	Edifícios construídos entre 2006 e 2011
n_edificios_estrut_betao	Edifícios com estrutura de betão armado
n_edificios_estrut_com_placa	Edifícios com estrutura de paredes de alvenaria com placa
n_edificios_estrut_sem_placa	Edifícios com estrutura de paredes de alvenaria sem placa

n_edificios_estrut_adobe_pedra	Edifícios com estrutura de paredes de adobe ou alvenaria de pedra solta
n_edificios_estrut_outra	Edifícios com outro tipo de estrutura
n_alojamentos	Total de Alojamentos
n_alojamentos_familiares	Alojamentos familiares
n_alojamentos_fam_classicos	Alojamentos familiares clássicos
n_alojamentos_fam_n_classicos	Alojamentos familiares não clássicos
n_alojamentos_colectivos	Alojamentos colectivos
n_classicos_res_habitual	Alojamentos clássicos de residência habitual
n_alojamentos_res_habitual	Alojamentos familiares de residência habitual
n_alojamentos_vagos	Alojamentos familiares vagos
n_res_habitual_com_agua	Alojamentos familiares de residência habitual com água
n_res_habitual_com_retrete	Alojamentos familiares de residência habitual com retrete
n_res_habitual_com_esgotos	Alojamentos familiares de residência habitual com esgotos
n_res_habitual_com_banho	Alojamentos familiares de residência habitual com banho
n_res_habitual_area_50	Alojamentos familiares clássicos de residencia habitual com área até 50 m2
n_res_habitual_area_50_100	Alojamentos familiares clássicos de residencia habitual com área de 50 m2 a 100 m2
n_res_habitual_area_100_200	Alojamentos familiares clássicos de residencia habitual com área de 100 m2 a 200 m2
n_res_habitual_area_200	Alojamentos familiares clássicos de residencia habitual com área maior que 200 m2
n_res_habitual_1_2_div	Alojamentos familiares clássicos de residência habitual com 1 ou 2 divisões
n_res_habitual_3_4_div	Alojamentos familiares clássicos de residência habitual com 3 ou 4 divisões
n_res_habitual_estac_1	Alojamentos familiares clássicos de residencia habitual com estacionamento p/ 1 veículo
n_res_habitual_estac_2	Alojamentos familiares clássicos de residencia habitual com estacionamento p/ 2 veículos
n_res_habitual_estac_3	Alojamentos familiares clássicos de residencia habitual com estacionamento p/ 3 ou + veículos
n_res_habitual_prop_ocup	Alojamentos familiares clássicos de residência habitual com proprietário ocupante
n_res_habitual_arrend	Alojamentos familiares clássicos de residência habitual arrendados
n_familias_classicas	Total de famílias clássicas
n_familias_institucionais	Total de famílias institucionais
n_familias_classicas_1ou2_pess	Famílias clássicas com 1 ou 2 pessoas

n_familias_classicas_3ou4_pess	Famílias clássicas com 3 ou 4 pessoas
n_familias_classicas_npes65	Famílias clássicas com pessoas com 65 ou mais anos
n_familias_classicas_npes14	Famílias clássicas com pessoas com menos de 15 anos
n_familias_classic_sem_desemp	Famílias clássicas sem desempregados
n_familias_classic_1desempreg	Famílias clássicas com 1 desempregado
n_familias_class_2mais_desemp	Famílias clássicas com + do que 1 desempregado
n_nucleos_familiares	Total de núcleos familiares residentes
n_nucleos_1filh_nao_casado	Núcleos com 1 filho não casado
n_nucleos_2filh_nao_casado	Núcleos com 2 filhos não casados
n_nucleos_filh_inf_6anos	Núcleos com filhos de idade inferior a 6 anos
n_nucleos_filh_inf_15anos	Núcleos c/ filhos c/ menos de 15 anos
n_nucleos_filh_mais_15anos	Núcleos c/ filhos todos c/ mais de 15 anos
n_individuos_present	Total de individuos presentes
n_individuos_present_h	Total de homens presentes
n_individuos_present_m	Total de mulheres presentes
n_individuos_resident	Total deindivíduos residentes
n_individuos_resident_h	Total de homens residentes
n_individuos_resident_m	Total de mulheres residentes
n_individuos_resident_0a4	Indivíduos residentes com idade entre 0 e 4 anos
n_individuos_resident_5a9	Indivíduos residentes com idade entre 5 e 9 anos
n_individuos_resident_10a13	Indivíduos residentes com idade entre 10 e 13 anos
n_individuos_resident_14a19	Indivíduos residentes com idade entre 14 e 19 anos
n_individuos_resident_15a19	Indivíduos residentes com idade entre 15 e 19 anos
n_individuos_resident_20a24	Indivíduos residentes com idade entre 20 e 24 anos
n_individuos_resident_20a64	Indivíduos residentes com idade entre 20 e 64 anos
n_individuos_resident_25a64	Indivíduos residentes com idade entre 25 e 64 anos
n_individuos_resident_65	Indivíduos residentes com idade superior a 64 anos
n_individuos_resident_h_0a4	Homens residentes com idade entre 0 e 4 anos

n_individuos_resident_h_5a9	Homens residentes com idade entre 5 e 9 anos
n_individuos_resident_h_10a13	Homens residentes com idade entre 10 e 13 anos
n_individuos_resident_h_14a19	Homens residentes com idade entre 14 e 19 anos
n_individuos_resident_h_15a19	Homens residentes com idade entre 15 e 19 anos
n_individuos_resident_h_20a24	Homens residentes com idade entre 20 e 24 anos
n_individuos_resident_h_20a64	Homens residentes com idade entre 20 e 64 anos
n_individuos_resident_h_25a64	Homens residentes com idade entre 25 e 64 anos
n_individuos_resident_h_65	Homens residentes com idade superior a 64 anos
n_individuos_resident_m_0a4	Mulheres residentes com idade entre 0 e 4 anos
n_individuos_resident_m_5a9	Mulheres residentes com idade entre 5 e 9 anos
n_individuos_resident_m_10a13	Mulheres residentes com idade entre 10 e 13 anos
n_individuos_resident_m_14a19	Mulheres residentes com idade entre 14 e 19 anos
n_individuos_resident_m_15a19	Mulheres residentes com idade entre 15 e 19 anos
n_individuos_resident_m_20a24	Mulheres residentes com idade entre 20 e 24 anos
n_individuos_resident_m_20a64	Mulheres residentes com idade entre 20 e 64 anos
n_individuos_resident_m_25a64	Mulheres residentes com idade entre 25 e 64 anos
n_individuos_resident_m_65	Mulheres residentes com idade superior a 64 anos
n_indiv_resident_n_ler_escrv	Indivíduos residentes sem saber ler nem escrever
n_ind_resident_fensino_1bas	Indivíduos residentes a frequentar o 1º ciclo do ensino básico
n_ind_resident_fensino_2bas	Indivíduos residentes a frequentar o 2º ciclo do ensino básico
n_ind_resident_fensino_3bas	Indivíduos residentes a frequentar o 3º ciclo do ensino básico
n_ind_resident_fensino_sec	Indivíduos residentes a frequentar o ensino secundário
n_ind_resident_fensino_possec	Individuos residentes a frequentar o ensino pós-secundário
n_ind_resident_fensino_sup	Indivíduos residentes a frequentar um curso superior
n_ind_resident_ensincomp_1bas	Indivíduos residentes com o 1º ciclo do ensino básico completo

n_ind_resident_ensinco mp_2bas	Indivíduos residentes com o 2º ciclo do ensino básico completo
n_ind_resident_ensinco mp_3bas	Indivíduos residentes com o 3º ciclo do ensino básico completo
n_ind_resident_ensinco mp_sec	Indivíduos residentes com o ensino secundrio completo
n_ind_resident_ensinco mp_posec	Individuos residentes com o ensino pós-secundário
n_ind_resident_ensinco mp_sup	Indivíduos residentes com um curso superior completo
n_ind_resid_desemp_pr oc_1emprg	Indivíduos residentes desempregados à procura do 1º emprego
n_ind_resid_desemp_pr oc_emprg	Indivíduos residentes desempregados à procura de novo emprego
n_ind_resid_empregado s	Indivíduos residentes empregados
n_ind_resid_pens_refor m	Indivíduos residentes pensionistas ou reformados
n_ind_resid_sem_act_ec on	Indivíduos residentes sem actividade económica
n_ind_resid_empreg_se ct_prim	Indivíduos residentes empregados no sector primário
n_ind_resid_empreg_se ct_seq	Indivíduos residentes empregados no sector secundário
n_ind_resid_empreg_se ct_terc	Indivíduos residentes empregados no sector terciário
n_ind_resid_estud_mun _resid	Indivíduos residentes a estudarem no municipio de residência
n_ind_resid_trab_mun_r esid	Indivíduos residentes a trabalharem no municipio de residência
Densidade_populacional	Nº de habitantes por unidade de área
1º quintil de rendimentos	1º quintil de rendimentos
2º quintil de rendimentos	2º quintil de rendimentos
3º quintil de rendimentos	3º quintil de rendimentos
4º quintil de rendimentos	4º quintil de rendimentos
5º quintil de rendimentos	5º quintil de rendimentos

ANEXO 4 – Aproximação das freguesias utilizadas

FREGUESIAS VALIDAÇÃO - ESCALA 1 : 500,00